

*Measurement  
and  
Evaluation  
in the  
Elementary School*



8973  
ARUP GHOSH

S. C. E. R. T

25/3 Ballygungee  
circular Road

Cal-19

Phn - 476 9119 8475 4377





***Measurement and Evaluation  
in the  
Elementary School***

# **Measurement and Evaluation in the Elementary School**

**Harry A. Greene, Ph.D.**

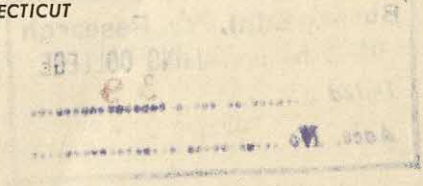
PROFESSOR OF EDUCATION AND  
DIRECTOR OF BUREAU OF EDUCATIONAL RESEARCH AND SERVICE  
UNIVERSITY OF IOWA

**Albert N. Jorgensen, Ph.D.**

PRESIDENT  
UNIVERSITY OF CONNECTICUT

**J. Raymond Gerberich, Ph.D.**

PROFESSOR OF EDUCATION AND  
DIRECTOR OF BUREAU OF EDUCATIONAL RESEARCH AND SERVICE  
UNIVERSITY OF CONNECTICUT



**LONGMANS, GREEN AND CO.**

NEW YORK LONDON TORONTO

1953

LONGMANS, GREEN AND CO., INC.  
55 FIFTH AVENUE, NEW YORK 3

LONGMANS, GREEN AND CO. LTD.  
6 & 7 CLIFFORD STREET, LONDON W 1

LONGMANS, GREEN AND CO.  
215 VICTORIA STREET, TORONTO 1

GREENE, JORGENSEN & GERBERICH  
MEASUREMENT AND EVALUATION IN THE  
ELEMENTARY SCHOOL

371.26

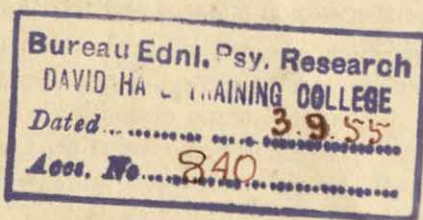
GRE

COPYRIGHT • 1953  
BY LONGMANS, GREEN AND CO., INC.

ALL RIGHTS RESERVED, INCLUDING THE RIGHT TO REPRODUCE  
THIS BOOK, OR ANY PORTION THEREOF, IN ANY FORM

PUBLISHED SIMULTANEOUSLY IN THE DOMINION OF CANADA BY  
LONGMANS, GREEN AND CO., TORONTO

SECOND EDITION



Printed in the United States of America  
Van Rees Press • New York



## ***Prefatory Statement***

PREVIOUS EDITIONS of this book have proven to be dependable and welcome instruments of instruction and guidance in the field of educational tests and measurements. For over twenty years many hundreds of young teachers have obtained their first grasp of the problems and possibilities of measurement and evaluation from the pages of the predecessors of this volume. But even a timely and successful professional book requires revision.

This book, as was true of its predecessor, is designed especially for the use of elementary-school teachers and students of elementary education. Essentially it is a completely revised treatment of the authors' earlier volume which appeared in 1942 under a similar title. It continues to present the practical introductory discussion of the essential principles of measurement and evaluation which students and teachers in general find readable and valuable. Certain recent and significant changes in points of view and in methods and techniques of measurement, as well as an obvious lack of timeliness in certain of the illustrative materials in the earlier edition, now serve to make a revision desirable. The continued interest of instructors and students in the type of treatment presented in the earlier volumes has served to encourage the authors in the preparation of this further revision.

The group specifically addressed in this volume is comprised of students and teachers whose major interests and responsibilities are in the elementary school. A second volume, parallel in general organi-

zation and treatment, is just as specifically addressed to those teachers and students who primarily face the problems of instruction, measurement, and evaluation at the secondary-school level. Illustrations, examples, and problems are chosen from material of suitable interest and concern to the reader. Many of the problems of measurement and evaluation are common to both the elementary and the secondary levels. The illustrations, however, are more meaningful if chosen from fields close to the fields of interest of the students and teachers. The present revision of these two volumes brings this treatment of measurement and evaluation quite up to the best thought and practices of 1953.

As has been true throughout the history of education, the decade just past has been marked by numerous highly significant developments in curricular points of view, in instructional methods and materials, as well as in measurement and evaluation techniques. A special effort has been made to retain the effective presentation of the whole problem of measurement and evaluation in good proportion and with the common-sense perspective of the earlier volumes. In addition an attempt has been made to broaden the point of view reflected in the earlier treatments and to introduce the student to easily comprehended discussion of the newest and best evaluative techniques that have thus far appeared. In this treatment, special emphasis is given to methods and materials designed for the measurement of intelligence and the evaluation of certain of the more intangible aspects of the child's personality. Many of the instruments and techniques presented here are so new in the field of educational measurement that only recently have they proved their dependable worth to the practical educator.

In this revision the authors have continued to place a heavy stress on the crucial and practical problems of improving all types of teacher-made examinations and tests. By principle and by example, the construction, improvement, use, and interpretation of all types of evaluative and measuring devices are treated in detail. Extensive new material is presented on the measurement of personality and on performance tests, evaluative tools and techniques, and graphical representation. The simplified treatment of the statistical problems of test interpretation presented in the earlier editions is continued in this volume. New problems of test interpretation closely related to the teacher's actual needs have been prepared. A new revision of the workbook designed to accompany this text is also in preparation.



This revised volume is planned to provide a complete and systematic handbook for any student or teacher requiring a straightforward and understandable discussion of all of the fundamental ideas and techniques of evaluation in the classroom. It is written from the point of view of the classroom teacher, and at all possible points non-technical language is used. In instances in which technical language cannot be avoided, such terms are introduced in context, and are defined and illustrated. Many words that may lie outside of the experience of the reader are included in the Glossary.

It is believed that classroom teachers, supervisors, and students in training for teaching and supervision will find in these pages a carefully written fundamental text on the principles of measurement and evaluation in education. It is hoped that a main contribution of this revision may be, not so much in the novel points of view or advances in the theory or technicalities of test construction as it is in the plainness of exposition and balance of treatment of many points which to some might otherwise seem over-technical. The authors themselves think of it as a first book in measurement and evaluation for those who at the time of studying it may know very little, if anything, about measurement in education and its application to the problems of improving classroom instruction. This volume offers carefully selected suggestions of ways in which measurement and evaluative instruments may be effectively used in the teaching of children. In addition, many general hints are given for the guidance of the student and teacher in constructing, selecting, using, and interpreting all types of educational tests as valuable aids in accomplishing this task.

Grateful acknowledgments are here expressed to the many experienced teachers and supervisors, as well as graduate students and colleagues, who have contributed directly or indirectly to improvements in the formulation and statement of much of the material incorporated into this volume. The authors are especially indebted to the many users of the earlier editions of this text who by their friendly and critical comments have stimulated and encouraged the development of this volume in its present form.

THE AUTHORS



# Contents

<b>I</b>	<b>MEASUREMENT, EVALUATION, AND THE CLASSROOM TEACHER</b>	<b>I</b>
1.	Need for measurement and evaluation in education	1
2.	Meaning of educational evaluation	3
3.	General characteristics of educational tests	5
4.	Types of tests to use	8
5.	Interpretation of test results	12
6.	Practical aspects of classroom measurement	12
7.	Organization of this book	16
<b>2</b>	<b>DEVELOPMENT OF EDUCATIONAL AND MENTAL MEASUREMENT</b>	<b>19</b>
1.	Measurement to 1800	20
2.	Educational testing from 1800 to 1900	22
3.	Educational measurement and evaluation from 1900 to the present	24
4.	Intelligence testing from 1800 to 1900	28
5.	Intelligence measurement from 1900 to the present	29
6.	Personality evaluation from 1800 to the present	32
7.	Present status of educational and mental measurement and evaluation	33

<b>3</b>	<b>EDUCATIONAL AND MENTAL MEASURING INSTRUMENTS AND TECHNIQUES . . . . .</b>	<b>37</b>
1.	General classification of tests . . . . .	37
2.	Educational tests . . . . .	44
3.	Intelligence tests . . . . .	54
4.	Personality inventories and evaluations . . . . .	58
<b>4</b>	<b>ESSENTIAL QUALITIES OF A GOOD MEASURING INSTRUMENT OR TECHNIQUE . . . . .</b>	<b>65</b>
1.	Validity . . . . .	66
2.	Reliability . . . . .	72
3.	Practicality . . . . .	79
4.	Comparability . . . . .	81
5.	Utility . . . . .	82
<b>5</b>	<b>CONSTRUCTING AND USING STANDARDIZED TESTS . . . . .</b>	<b>86</b>
1.	Constructing standardized tests . . . . .	86
2.	Practical uses of standardized tests . . . . .	105
3.	Planning testing programs . . . . .	119
4.	Selecting tests . . . . .	123
5.	Administering tests . . . . .	126
6.	Scoring tests . . . . .	128
7.	Analyzing and interpreting results of testing . . . . .	134
<b>6</b>	<b>CONSTRUCTING AND USING ORAL AND ESSAY TESTS . . . . .</b>	<b>138</b>
1.	Classroom testing . . . . .	138
2.	Oral examinations . . . . .	139
3.	Essay examinations . . . . .	141
4.	Improving essay examinations . . . . .	152
<b>7</b>	<b>CONSTRUCTING AND USING INFORMAL OBJECTIVE TESTS . . . . .</b>	<b>160</b>
1.	Characteristics of classroom testing . . . . .	161
2.	Advantages and limitations of informal objective tests . . . . .	163

Consider of  
good test

3.	Construction and use of informal objective tests . . . . .	167
4.	Types of objective items . . . . .	177
5.	Constructing objective test items . . . . .	186
6.	Using results of informal objective testing . . . . .	194

## 8    CONSTRUCTING AND USING PERFORMANCE TESTS . . . . . 199

1.	Nature of performance tests . . . . .	200
2.	Object tests . . . . .	202
3.	Performance measures . . . . .	204
4.	Product evaluation . . . . .	207
5.	Constructing performance tests . . . . .	213
6.	Using results of performance testing . . . . .	215

## 9    CONSTRUCTING AND USING EVALUATION TOOLS AND TECHNIQUES . . . . . 217

1.	Meaning of evaluation . . . . .	218
2.	Evaluative tests . . . . .	219
3.	Other evaluative tools . . . . .	225
4.	Evaluative techniques . . . . .	232
5.	Evaluative tools and techniques in the classroom . . . . .	234

## 10   USING INTELLIGENCE AND APTITUDE TESTS . . . . . 238

1.	Nature of intelligence . . . . .	239
2.	Measurement of intelligence . . . . .	242
3.	General intelligence tests . . . . .	244
4.	Specific intelligence tests . . . . .	250
5.	Group-factor tests of intelligence . . . . .	253
6.	Performance tests of intelligence and aptitude . . . . .	256
7.	Derived results of intelligence testing . . . . .	257
8.	Distribution of intelligence . . . . .	265
9.	Derived measures relating intelligence and achievement . . . . .	266
10.	General procedures for intelligence testing . . . . .	268
11.	Values and uses of different types of tests . . . . .	270



<b>11</b>	<b>USING PERSONALITY INSTRUMENTS AND TECHNIQUES</b>	278
1.	Nature of personality	279
2.	Techniques of personality measurement	281
3.	Measurement of attitudes	284
4.	Measurement of interests	287
5.	Measurement of emotional adjustment	291
6.	Evaluative techniques	297
7.	Measurement of total personality	303
<b>12</b>	<b>SUMMARIZING THE RESULTS OF MEASUREMENT</b>	308
1.	Classification and tabulation of test scores	309
2.	Measures of central tendency	317
3.	Measures of variability	328
<b>13</b>	<b>INTERPRETING THE RESULTS OF MEASUREMENT</b>	339
1.	Test scores	339
2.	Formal types of derived scores	342
3.	Informal types of derived scores	350
4.	Graphical representation	355
5.	Major types of norms	363
<b>14</b>	<b>DETERMINING RELATIONSHIPS AMONG THE RESULTS OF MEASUREMENT</b>	371
1.	Relationship between sets of test scores	371
2.	Computation of Pearson product-moment correlation coefficient	374
3.	Meaning of correlation coefficients	381
4.	Practical uses of correlation coefficients	385
<b>15</b>	<b>MEASURING AND EVALUATING IN THE RECEPTIVE LANGUAGE ARTS</b>	392
1.	Importance of listening and reading as receptive language skills	393
2.	Identification of factors affecting listening and reading	399

3. Determination of reading readiness . . . . .	402
4. Measuring oral reading and listening comprehension . . . . .	405
5. Analysis and diagnosis in silent reading . . . . .	408
6. Corrective exercises in reading . . . . .	411

## 16 MEASURING AND EVALUATING IN THE EXPRESSIVE LANGUAGE ARTS . . . . . 419

1. Identification of language abilities . . . . .	419
2. Measurement and diagnosis of language abilities . . . . .	425
3. Remedial instruction in language . . . . .	434
4. Measurement and remediation in spelling . . . . .	439
5. Measurement and remediation in handwriting . . . . .	448

## 17 MEASURING AND EVALUATING IN THE SOCIAL STUDIES . . . . . 462

1. Scope of social studies . . . . .	463
2. Outcomes of social studies . . . . .	465
3. Standardized social studies tests . . . . .	467
4. Classroom testing and evaluating in social studies . . . . .	475
5. Corrective work in social studies . . . . .	477

## 18 MEASURING AND EVALUATING IN ELEMENTARY-SCHOOL MATHEMATICS . . . . . 481

1. Course content and organization in arithmetic . . . . .	482
2. Measurement of general achievement in arithmetic . . . . .	488
3. Diagnostic testing in arithmetic skills . . . . .	491
4. Testing problem-solving ability . . . . .	493
5. Remedial instruction in arithmetic . . . . .	494
6. Prediction of success in later mathematics . . . . .	500

## 19 MEASURING AND EVALUATING IN THE ELEMENTARY SCIENCES . . . . . 504

1. Scope of elementary sciences . . . . .	505
2. Measurement in elementary sciences . . . . .	509
3. Standardized tests in elementary sciences . . . . .	511
4. Testing in elementary sciences . . . . .	515



5. Diagnosis and remedial teaching in elementary sciences . . . . .	522
---	-----

## 20 MEASURING AND EVALUATING IN THE FINE ARTS . . . . . 526

1. Measurable qualities in music . . . . .	527
2. Measurement of musical talent . . . . .	530
3. Measurement of musical achievement . . . . .	533
4. Characteristics and aims of art education . . . . .	537
5. Measurement of art abilities and achievement . . . . .	539

## 21 MEASURING AND EVALUATING IN HEALTH AND PHYSICAL EDUCATION . . . . . 546

1. Scope and aims of health education . . . . .	546
2. Measurement and evaluation in health education . . . . .	548
3. Prevention and diagnosis in health education . . . . .	552
4. Objectives of physical education . . . . .	553
5. Measurement in physical education . . . . .	554
6. Diagnosis in physical education . . . . .	560

## 22 MEASURING AND EVALUATING GENERAL EDUCATIONAL ACHIEVEMENT . . . . . 564

1. Measurement of general achievement . . . . .	564
2. General achievement batteries . . . . .	567
3. Achievement batteries in skill areas . . . . .	574
4. Achievement batteries in content areas . . . . .	578
5. Specialized batteries . . . . .	579

Glossary . . . . .	583
--------------------	-----

Appendix. Publishers of tests illustrated and discussed . . . . .	599
---	-----

Index of Names . . . . .	601
--------------------------	-----

Index of Subject Matter . . . . .	607
-----------------------------------	-----



## Tables

1. Scores assigned by ten teachers to an essay and a true-false examination over the same material in geography . . . .	78
2. Discriminative power of test items in percentages of success by superior and inferior groups . . . . .	92
3. Grade equivalents corresponding to each subtest standard score and the median standard score for the Iowa Language Abilities Test . . . . .	97
4. Age-at-grade norms for the total language score of the Iowa Basic Skills Tests . . . . .	99
5. Percentile norms for school averages on the total language score of the Iowa Basic Skills Tests . . . . .	101
6. Distribution of mental ability in a ninth-grade class in terms of average grade placement . . . . .	107
7. Shifting standards of expectancy . . . . .	145
8. Distribution of intelligence quotients in a normal population . . . . .	265
9. Reading test scores of 37 ninth-grade pupils in alphabetical order of last names . . . . .	310
10. Reading test scores of 37 ninth-grade pupils in descending order . . . . .	310
11. Reading test scores of 37 ninth-grade pupils in frequency distributions . . . . .	311

12.	Relation between range of scores and size of class intervals	313
13.	Reading test scores of 37 ninth-grade pupils in a grouped frequency distribution	314
14.	Computation of the arithmetic mean for the grouped frequency distribution of 37 reading test scores	320
15.	Computation of the median for the grouped frequency distribution of 37 reading test scores	325
16.	Data showing identical means but unlike variability	329
17.	Computation of the standard deviation for ungrouped data	333
18.	Computation of the standard deviation for the grouped frequency distribution of 37 reading test scores	334
19.	Computation of deciles and percentiles for the grouped frequency distribution of 37 reading test scores	348
20.	Assignment of relative ranks to arithmetic test scores	351
21.	Suggested weightings for marks in obtaining composite scores	354
22.	Class frequencies, cumulative frequencies, and cumulative relative frequencies for 37 reading test scores	356
23.	Mental age norms for the Pintner General Ability Tests, Verbal Series	364
24.	Grade placement and age norms for the California Language Test	365
25.	Percentile norms for the Cooperative Mechanics of Expression Test, A	366
26.	Percentile norms for the Aspects of Personality Inventory	367
27.	Pairs of test scores	374
28.	Computation of the Pearson product-moment coefficient of correlation between speed and comprehension scores on a certain reading test	376
29.	Moments of cells in a double-entry table	378
30.	Percentages of forecasting efficiency for certain values of $r$	383
31.	Diagnostic and remedial chart: Language usage and punctuation	436
32.	Remedial chart: Spelling	446
33.	Handwriting quality and speed standards	450
34.	Tentative norms for writing a practice sentence	457

35. Possible scope of drill units in whole numbers . . . . .	496
36. Analysis of problem-solving . . . . .	497
37. Summary of tests and timing: Stanford Achievement Tests	568
38. Summary of tests and timing: Metropolitan Achievement Tests . . . . .	569
39. Summary of tests and timing: Coordinated Scales of Attainment . . . . .	570
40. Summary of parts and timing: Modern School Achievement Tests . . . . .	571
41. Summary of tests and timing: Gray-Votaw-Rogers General Achievement Tests . . . . .	571
42. Summary of tests and timing: Master Achievement Tests	572
43. Summary of tests and timing: National Achievement Tests	573
44. Summary of tests and timing: Cooperative Achievement Tests for the junior high school . . . . .	574
45. Summary of tests and timing: Iowa Every-Pupil Tests of Basic Skills . . . . .	575
46. Summary of tests and timing: California Achievement Tests . . . . .	576
47. Summary of tests and timing: American School Achievement Tests . . . . .	577
48. Summary of tests and timing: Progressive Tests in Social and Related Sciences . . . . .	578



## Figures

1. Contrast between analysis and diagnosis . . . . .	50
2. The principle of sampling . . . . .	76
3. Diagnostic profile chart for the California Reading Test . . . . .	108
4. Individual educational chart for the Gray-Votaw-Rogers General Achievement Test . . . . .	110
5. Cutout scoring stencil for the Iowa Every-Pupil Tests of Basic Skills . . . . .	130
6. International Test Scoring Machine . . . . .	132
7. Samples of machine-scored answer sheets . . . . .	133
8. Effect of limited sampling on test scores . . . . .	143
9. Marks assigned to an English examination by 142 teachers . . . . .	144
10. Effect of extensive sampling on test scores . . . . .	164
11. Check list of student reactions in finding an object under a microscope . . . . .	205
12. Handwriting scale of the California Achievement Tests . . . . .	208
13. Working drawing of a wood block . . . . .	211
14. Diagram of pieces of a dress pattern . . . . .	213
15. Sample profile chart for the California Arithmetic Test . . . . .	226
16. Sample graphic record of pupil progress for the Metropolitan Achievement Tests . . . . .	227
17. Sample American Council on Education Cumulative Rec- ord for Elementary and Secondary Schools . . . . .	230
18. Sample class analysis chart for the Metropolitan Achieve- ment Tests . . . . .	232

19. Tests of the Pintner-Paterson "Long" Performance Scale .	256
20. Percentages of persons in a normal population at different levels of intelligence . . . . .	266
21. Sample sociogram . . . . .	301
22. Behavioral categories and their major relations . . . . .	302
23. Midpoints and limits of a class interval . . . . .	313
24. Moments of force and the arithmetic mean . . . . .	318
25. Moments of force for the 37 reading test scores . . . . .	321
26. Assumptions concerning the distribution of scores in class intervals in the computation of the arithmetic mean and median . . . . .	327
27. Measures of variability for homogeneous and heterogeneous data . . . . .	329
28. Arithmetic mean and standard deviation . . . . .	331
29. Frequency polygon of 37 reading test scores . . . . .	357
30. Histogram of 37 reading test scores . . . . .	359
31. Cumulative frequency graph of 37 reading test scores . . . . .	361

## Exercises

### TABULATING TEST SCORES

1. Tabulating arithmetic test scores . . . . . 317
2. Tabulating language test scores . . . . . 317
3. Tabulating spelling test scores . . . . . 317

### COMPUTING THE ARITHMETIC MEAN

4. Computing the arithmetic mean of arithmetic test scores 323
5. Computing the arithmetic mean of language test scores . 323
6. Computing the arithmetic mean of spelling test scores . 323

### COMPUTING THE MID-MEASURE AND MEDIAN

7. Finding the mid-measure of arithmetic test scores . . 326
8. Finding the mid-measure of language test scores . . . 326
9. Computing the median of arithmetic test scores . . . 326
10. Computing the median of language test scores . . . 326

### COMPUTING THE STANDARD DEVIATION

11. Computing the standard deviation of arithmetic test scores . . . . . 337
12. Computing the standard deviation of language test scores . . . . . 337



**COMPUTING DERIVED SCORES**

- |  |     |
|--|-----|
| 13. Assigning <i>T</i> -scores to certain arithmetic test scores . . . | 355 |
| 14. Computing certain percentiles for arithmetic test scores . . .     | 355 |
| 15. Assigning relative ranks to spelling test scores . . .             | 355 |
| 16. Establishing marks for arithmetic test scores . . .                | 355 |
| 17. Obtaining a weighted composite score for a pupil . . .             | 355 |

**CONSTRUCTING GRAPHS**

- |  |     |
|--|-----|
| 18. Constructing a frequency polygon for arithmetic test scores . . . . .          | 363 |
| 19. Constructing a histogram for language test scores . . .                        | 363 |
| 20. Constructing a cumulative frequency graph for arithmetic test scores . . . . . | 363 |

**ESTIMATING PERCENTILES AND PERCENTILE RANKS**

- |  |     |
|--|-----|
| 21. Estimating certain percentiles for arithmetic test scores . . .          | 363 |
| 22. Estimating certain percentile ranks for arithmetic test scores . . . . . | 363 |

**USING TEST NORMS**

- |   |     |
|---|-----|
| 23. Obtaining mental ages and intelligence quotients from intelligence test norms . . . . . | 368 |
| 24. Obtaining grade and age equivalents from achievement test norms . . . . .               | 368 |
| 25. Obtaining percentile ranks from achievement test norms . . .                            | 368 |
| 26. Obtaining percentile ranks from personality inventory norms . . . . .                   | 369 |

**COMPUTING CORRELATION COEFFICIENTS AND ESTIMATING RELIABILITY COEFFICIENTS**

- |   |     |
|---|-----|
| 27. Computing the correlation coefficient between reading test and vocabulary test scores . . . . . | 390 |
| 28. Estimating the reliability coefficient of a vocabulary test . . .                               | 390 |
| 29. Estimating the reliability coefficient of a history test . . .                                  | 390 |
| 30. Computing the standard error of measurement for a vocabulary test . . . . .                     | 390 |

## ***Measurement, Evaluation, and the Classroom Teacher***

THE PURPOSE of this chapter is to introduce the reader to the following general notions underlying educational measurement and evaluation, and to provide a preview of the contents and organization of this book:

- A. Measurement in education not a new idea.
- B. Urgent need for improved measurement and evaluation techniques.
- C. Importance of evaluation and measurement to the school and to the teacher.
- D. Characteristics of educational tests.
- E. Purposes tests do and do not serve.
- F. General problems of measurement and evaluation.
- G. Organization of this volume.

### **1 NEED FOR MEASUREMENT AND EVALUATION IN EDUCATION**

#### **Educational measurement not a new idea**

Teachers have always endeavored to measure the results of their teaching efforts as indicated by the progress of their pupils toward desired educational goals. Many have been equally concerned about the need to diagnose and remedy revealed defects in instruction. However, only recently has any large degree of accuracy been injected into their methods of measurement and diagnosis. Measure-



ment of progress, evaluation of efficiency of instruction, and the accompanying attempts at diagnosis were largely a matter of personal observation and judgment on the part of the individual teacher as recently as two decades ago. Actually, the recent development of modern educational instruments of measurement and evaluation may be regarded as an extension and improvement of an old practice. The modern educational measuring instrument presents a surprisingly accurate picture of the course objectives as well as an analysis of the underlying skills, knowledges, concepts, understandings, and other outcomes upon which accomplishment in different subject-matter fields depends. It points out weaknesses in learning and instructional procedures. It permits the establishment of specific and objective goals of achievement that are based upon the actual attainments of children under typical school conditions. Educational tests and the information resulting from their use in the classroom have come to be almost universally identified with good teaching practice. Today the professionally equipped teacher is expected to be well versed in their construction, selection, and use in the classroom.

### Early recognition of the need for tests in the classroom

For many years the teacher's estimate was accepted as the sole measure of a pupil's ability or accomplishment. Studies of the reliability of such methods gradually cast serious doubts on their accuracy, with the result that since that time a continuous search for more dependable measures has been carried on. Today the testing movement has passed through the first stages of its development. Thirty years ago it was necessary to popularize the idea. Now the advantages of standardized and informal objective tests are recognized by most educators and by many laymen. Moreover, the tests themselves have been greatly improved in content and in structure as a result of the critical analysis and refinement to which they have been subjected. The most enthusiastic students of educational tests are and should be their own severest critics. The specific shortcomings of tests are coming to be fully realized. They are not mysterious instruments for the confusion of the uninitiated, but are useful devices for assisting the professionally minded educator to improve the conditions under which children learn and teachers teach. If they aid in the accomplishment of this, the primary purpose of all



school supervision, and for that matter of all of the educative process, they are thoroughly justified.

## 2 MEANING OF EDUCATIONAL EVALUATION

Several different attitudes toward the use of educational measurements in the school have held sway at various times since the objective approach to the measurement of pupil intelligence and achievement made its appearance shortly after the beginning of the twentieth century. These different attitudes or outlooks may be called by the following names: (1) testing, (2) measuring, and (3) evaluating and appraising.

The first concept chronologically was that of testing, which considered the development of objective devices for testing intelligence and achievement of pupils to be of major importance. This attitude was doubtless the result of the early need for the development of objective instruments, for such instruments were not available in any significant quantity for some years after the concept of objectivity of tests first made its appearance in the field of education.

When objective tests became fairly numerous and classroom teachers began to use objective methods in their own examinations, attention turned more toward the use of test results and toward the development of instruments for measuring certain of the more elusive types of instructional outcomes that do not lend themselves readily to objective measurement. This period may be characterized as one during which the major approach was that of measuring.

The quite recent development of the evaluation and appraisal concept was doubtless impelled by the increasing realization that paper-and-pencil tests can measure only a limited portion of the outcomes of instruction and types of pupil behavior about which the teacher and other school officers need information. Therefore, the present view is that objective tests constitute probably the major type of evaluative instruments but that such other means of measurement as the anecdotal record, the interview, the questionnaire, the rating scale, and such tools as the individual pupil profile, the class record, the cumulative record, and the case study have a very significant place in the evaluation of pupil behavior and achievement. The evaluation concept has also doubtless been stimulated by the recent attention of educators and psychologists to the whole child and his behavior. This tendency to consider the child as a

whole, rather than as an individual whose behavior and abilities can be catalogued into a number of different compartments, places a definite responsibility on the user of tests and other instruments of evaluation for considering the child in this broad sense. It is through the application of the evaluation concept rather than of the narrower concepts of measuring and testing that this result is most effectively obtained.

Perhaps a more exact idea of the meaning of educational tests to the classroom teacher may be obtained most readily by considering the characteristics that distinguish them from other types of measuring instruments in education. In the first place, the educational test of standardized or semi-standardized form is more limited in its usefulness than the informal objective examination. This is necessarily so, since the standardized test must confine itself to the general aspects of the subject that can be covered by all classes, while the typical examination or objective test prepared by the classroom teacher covers a specific selection of the content or activities from the teacher's own class. In the second place, the items comprising standardized educational tests are commonly constructed and arranged in accordance with certain statistical and educational principles designed to produce more accurate measuring instruments. In the third place, the more useful standardized educational tests have been taken by a large sampling of school children under controlled conditions. From this use of the tests the norms that give meaning to test results and permit the interpretation of test scores are derived. In the fourth place, the more carefully constructed and valuable educational tests yield results that point the teacher's way to the application of specific remedial methods where needed.

Both informal objective tests and standardized educational tests are characterized by other features, such as validity, reliability, and objectivity. Validity refers to the truth of the picture of the ability or achievement revealed by the test. Reliability refers to the consistency with which the test reveals this picture. Objectivity refers to the extent to which the test results are affected by the personal judgment of the user. There are other important characteristics of educational tests, but these represent the major ones on which their meaning depends.



### 3 GENERAL CHARACTERISTICS OF EDUCATIONAL TESTS

#### Measurement in education

In most fields of human endeavor the most efficient results are obtained when the worker has clearly defined goals toward which to work and dependable instruments for determining progress. In fact, many critical workers in measurements are seriously questioning the defensibility of curricular objectives that are so vague or general as to defy measurement. A definite aim enables the worker to direct his efforts toward the particular task to be accomplished. By the proper use of instruments for measuring results it is possible for the worker to know what he has accomplished. Thus a reliable and analytic silent reading test will give to the teacher a measure of his relative success in developing silent reading skills in his class. Accurate measuring instruments also aid in discovering when emphasis has been misplaced. For example, a pupil who, in his elementary-school work, has been given an unusual emphasis on oral reading may satisfactorily pronounce words appearing on the printed page, but he may be sadly lacking in ability to get meaning from these same words.

Measuring instruments also make it possible for the worker to resort to experimental methods and thus to learn definitely whether materials and methods are effective. This is as true in the field of teaching as in other fields. Without specific aims the teacher cannot plan his work effectively. He cannot know, except in an indefinite way, what he is to do. A teacher without specific aims is like a person who starts out to walk to a certain place without any idea of which direction he is to take or how far away his destination may be. If, on the other hand, the goals of instruction are clear-cut and accurate, means of determining progress are provided and the probability of a timely arrival at the goal is greatly increased.

#### What educational tests are

Modern standardized and informal objective educational tests differ in several respects from the typical teacher-made examination of the discussion type. In the first place, the exercises used in educational tests are often much more carefully selected to coincide with



the purpose for which the test is designed than is true of questions in the ordinary essay examination. For example, an objective or a standardized test having for its purpose the measurement of ability to locate the states of the United States contains only such items as relate specifically to that purpose. The traditional essay-type examination frequently contains questions sampling into many different fields. In the second place, the items in a carefully made educational test are commonly arranged in accordance with certain principles of test construction so as to form an accurate measuring instrument. For the present it is enough for the reader to note that there are important principles of arrangement of items within a test that should be considered.

In the third place, the standardized test or scale is given to a large number of children of varying age and school classification, and from these results the norms are obtained. A fourth point of great importance, but only recently recognized and applied in connection with test development, is the fact that test scores yielded by narrow-function tests of the unit and analytic types are more readily translatable into specific remedial procedures. This means that a really valuable test, in addition to giving a cross-section of the instructional situation, must break down and identify the underlying skills so specifically that the results may be readily interpreted in terms of the specific kind and approximate amount of remedial attention needed.

Although the above discussion relates particularly to standardized tests, it should be recognized that the teacher can construct his own objective tests to serve purposes for which no suitable standardized tests are available or for which teacher-made tests are more satisfactory. The standardized test and the teacher-made test supplement each other, and both are important in a well-rounded testing program. Certain other measurement devices of a non-test type also have great significance in the evaluation of child behavior and of the results of classroom instruction.

### What objective tests do

Every alert teacher has at some time earnestly desired to know whether his instruction in certain school subjects was particularly superior or inferior, effective or ineffective. Many teachers have taken steps to answer this highly important question through the use

of one kind of measuring instrument or another, yet relatively few classroom teachers are utilizing to the fullest extent one of the most valuable instruments at their command. Supervisors and administrators are frequently obliged to admit that they do not have adequate data on which to base their decisions about the efficiency of a certain method of instruction, but must be guided largely by their own personal opinions. With the development of adequately valid and reliable measuring devices, useful objective information on such questions has become available.

Objective tests, either informal or standardized, are not panaceas for all educational inadequacies, but unquestionably the scores they afford are useful in the evaluation of instruction. Scores from standardized tests are objective and are given meaning by the process of standardization. For example, a quality score of 46 assigned to a certain handwriting sample is meaningless until it is understood that 46 points is standard quality for a fourth-grade child. If such a sample were scored as a second-grade product, it might be assigned a superior mark. If scored by an eighth-grade teacher, it might readily be given an inferior mark. Thus, by the use of standardized test scores, specific goals of achievement may be set up, and progress may be measured. Test norms themselves provide the basis for the more objective grade placement of pupils. Analytic and diagnostic tests make possible the discovery of pupils needing special corrective instruction. Vague objectives in the course of study may be pointed out and methods of instruction may be evaluated through the use of educational tests and critical interpretation of their results.

Experience in the use of tests in many school systems leads to the conclusion that quite often there is an intangible psychological effect resulting from the administration of a series of tests in a school. The experience of the children while taking the test, and the feeling on the part of the teacher that his work is being carefully checked, are both motivating forces making for better and more effective teaching and learning situations. This is, however, only a by-product of the use of the test and is no reason for allowing the work to stop short of a really constructive supervisory program.

It should be recognized that educational tests are incapable in and of themselves of directly improving instruction in any subject. They merely reveal the situation. In a sense a test may be thought of as an educational barometer. It reveals the educational atmospheric pressure, but does not do anything about it. Perhaps the parallel is



closer than at first appears. Just as low barometric readings indicate low atmospheric pressure and forecast changing weather conditions or storms, low achievement test scores may presage an unsatisfactory educational situation. As high or rising barometric records indicate fair weather, so high scores on educational tests indicate a satisfactory instructional situation.

The chief service of tests lies in their power to reveal the strengths and weaknesses of individual pupils or of the class as a whole. The use of tests must be followed by the next logical step, the development of a constructive supervisory program. It is not enough that weaknesses be revealed. They must be corrected by the use of properly constructed remedial exercises.

#### 4 TYPES OF TESTS TO USE

##### Significance attached to tests and examinations

It requires but a casual inspection of educational practices to discover the significance that is attached to tests and examinations by the school, as well as by the teacher, the pupil, and the parent. Pupils spend a great deal of time in preparing for and writing examinations. The school spends considerable time and money setting up an organization for the preparation and administration of examinations. Teachers devote much effort to the preparation, scoring, and marking of examination papers, while parents in general set far too much store by the marks earned by their children on school examinations.

Examinations play an important part in the public relations contacts of the school. To a certain extent they carry to the parents in the community the educational purposes of the school, the aims of specific subjects and courses, and the various emphases held important by the instructional agents of their school. Examinations in part serve as a means of revealing to both parent and pupil the basis for a pupil's scholastic rating, his promotions, failures, conditions, awards, and preparation for further educational work.

For the teacher, examinations focus attention on specific objectives and provide a means of determining his efficiency in achieving them. They aid in revealing overemphasis or wrong emphasis in teaching method and make possible the experimental evaluation of subject-matter organization. The very real value of the properly

constructed test or examination as an important teaching instrument is reflected in the recent tendency to recognize the special significance of low scores as well as high scores on tests and examinations. On the assumption that the basic objectives of the course are represented adequately in the item content of the test, a high score may reflect effective teaching, an accurate memory on the part of the pupil, or actual mastery of the essentials of the content. A low score, on the other hand, may be of greater significance to the pupil and to the instructor, since it identifies specifically the failure of the pupil to master concepts which the instructor considered of sufficient importance to include in the course and in the test. In this way the instructional weaknesses and the remedial needs of the individuals in the class are brought into sharp focus.

### **Teacher-made measures of achievement**

The emphasis on the use of the standardized test in most discussions of measurement problems often leads to the mistaken idea on the part of the student that these more formal types of tests are the most important measures of achievement. In most subject fields this is distinctly not the case. The use of some form of testing procedure for instructional purposes probably constitutes nine-tenths of the teacher's measurement activity in the classroom. Accordingly, much more attention should be given to the improvement of the teacher's informal measures of achievement.

### **Using informal objective tests in the classroom**

The informal objective examination has increased in popularity with great rapidity during the last two decades, although it is well recognized that there are certain areas of educational accomplishment in which it does not measure adequately. The successful construction of objective examinations calls for the application of many of the same principles of test construction as are involved in the development of standardized tests.

### **Uses of standardized educational tests**

Educational tests, because of their definiteness and objectivity, reveal to the teacher the status of the achievement of his class. They



point out individual pupil differences in capacity and achievement. If standardized, they set up specific goals of achievement for the teacher. They reveal the results of special types of emphasis, or of special methods of instruction. They open to the administrator and the teacher hitherto untouched sources of information useful in giving the pupil proper educational and vocational guidance. In their modern conception, they reveal to the teacher the specific weakness of individual pupils so definitely that he is in a position to apply effective instructional and corrective methods. Tests themselves have little or no power to bring about changes in pupil achievement as a mere result of their use. Their chief service is their power to reveal pupil strengths or weaknesses. The correction of weaknesses is another aspect of the supervisory problem.

### Using standardized tests in the classroom

Teachers themselves must assume a larger share of the responsibility for the use of educational tests in the classroom, and for the interpretation and application of the results after the tests have been given. Only by so doing does the teacher receive an adequate return from the use of tests. If this responsibility is to be wisely assumed, the teacher must have an understanding of the possibilities and the weaknesses of tests. He must be trained in their use and the interpretation of their results. He must be willing to exchange a certain amount of personal effort for the information concerning his teaching problems that the tests can furnish him.

Training in the use of tests comes as a result of their use. Opportunity for this training may be afforded through the preparation and use of informal objective tests as substitutes for the traditional examination, or it may be provided by undertaking a study of some supervisory problems of importance to the teacher in which standardized tests are used.

*Main uses of tests.* Three main types of uses of educational tests are noted, each resulting in a different point of view regarding the teacher's responsibility. Tests of a detailed diagnostic type designed to give the teacher precise information concerning the abilities and limitations of his pupils are instructional in their function. The responsibility for the use of such material should be the teacher's. Tests designed to be used more particularly for survey or supervisory purposes should probably be administered by persons other

than the teacher. The use of tests for administrative purposes, such as for pupil classification, gradation, or sectioning, may well be the joint responsibility of the teacher and the administrator. In other words, the function the tests are intended to perform determines where the responsibility for their use and interpretation lies.

*Selection of the test to use.* The criteria for tests—validity, reliability, adequacy, objectivity, practicality, administrability, scorability, economy, comparability, and utility—afford the teacher a tangible basis for the selection of the test to use for a certain purpose. In a general way, however, the teacher should depend upon the advice of persons who have made a special study of the tests, rather than attempt to apply these criteria personally. If affirmative answers to each of the following questions are available concerning a specific test, the teacher may feel reasonably safe in selecting it for use.

1. Does this particular test measure the skills, knowledges, concepts, understandings, applications, or appreciations I wish to measure?
2. How much time does it take to give the test? Is it long enough to give a reliable and consistent measure?
3. Is it easily and accurately scored?
4. Has it been widely used elsewhere?
5. Does it furnish accurate and extensive norms for comparison and interpretation?
6. Is the interpretation of the scores simple and clear?
7. Do the results point the way to a remedial program?
8. Is the test economical in terms of time and money cost per unit of reliable information furnished by it?

*Administration of the test.* One of the distinctive features of standardized tests is that they must be given under conditions closely approximating those under which they were standardized, if the results are to be meaningful. Accordingly the teacher should follow the directions furnished with such tests.

The attitude and the personality of the examiner are also important in the administration of a test. The whole purpose of the test is defeated if an unnatural response is obtained. In the giving of tests the greatest care must be exercised to secure the cheerful confidence of the pupils.

*Scoring the test papers.* Although the use of machine-scored tests and test-scoring services is growing rapidly, many users of educational tests feel that much of the value arising from their use in the



classroom is lost if the teacher himself does not have some first-hand contact with the papers. For the majority of the better tests, the task of scoring the papers is greatly simplified and objectified by the use of answer keys, scoring stencils, and in many cases mechanical devices. Whether the tests are hand-scored by teachers or mechanically scored as a part of a testing service, the answer sheets or the detailed records of the pupil's results should be made available to the teacher for individual instructional guidance.

## 5 INTERPRETATION OF TEST RESULTS

### Summarizing and interpreting the results of testing

Skill in summarizing and interpreting test results is dependent upon the mastery of the following statistical techniques:

1. A knowledge of why and how to classify and tabulate data.
2. A knowledge of how to find the common measures of central tendency.
3. A knowledge of how to express the variability of data.
4. A knowledge of how to determine the relationship between two or more groups of data.
5. A knowledge of how to derive and use norms and derived scores for purposes of comparison and interpretation of test results.
6. A knowledge of how to treat data for simple graphic presentation.

In addition to the use the classroom teacher may make of these skills in the proper interpretation and utilization of test scores, there is the application that may be made of them in the study of current educational literature. Reports of progress in education are filled with statistical terms and techniques. The teacher can scarcely hope to keep abreast of the times in his profession if he is unable to read current educational literature understandingly.

## 6 PRACTICAL ASPECTS OF CLASSROOM MEASUREMENT

### Diagnostic testing and remedial teaching

The analysis, identification, and measurement of the abilities that underlie and condition educational achievement unquestionably constitute the high point of the use of tests in educational practice. Forming the background of practically all possibilities of learning is that curiously interwoven maze of traits, tendencies, and predis-

positions known as mental ability. Naturally enough, instruments designed to sample into this field constitute an important unit of the teacher's diagnostic equipment.

*Intelligence.* The acceptance of the definition of intelligence as the capacity or power of the individual to learn or to adapt himself to new situations makes it relatively easy to set up devices for its measurement and interpretation. Intelligence tests are incapable of securing a direct measure of capacity unaffected by experience and training. They measure neither the actual process of learning, nor the quality of the learning equipment directly, but they provide the basis for inferences about the equipment from the amount of learning that has taken place under certain conditions. The value of the intelligence test lies in the opportunity it affords for making this inference quickly and on a reasonably objective basis. Thus the intelligence test, carefully used and critically interpreted, constitutes a most effective and useful instrument for classroom diagnosis. Not only do intelligence test scores provide valuable evidence of basic or general limitations and superiorities, but the related aptitude and group-factor tests offer most helpful hints about the existence of more highly specialized abilities or disabilities. In the last analysis, predictive tests that render such important service in certain types of educational and vocational guidance are specialized tests of intelligence.

*Personality.* In the sense that an individual's personality is revealed in all his behavior, this aspect of classroom measurement is all-inclusive. In a somewhat narrower sense, personality has to do with such forms of behavior as attitudes, interests, and emotional adjustment, all of which are important considerations in the classroom. Personality inventories and scales are doubtless still in their early stages of development, but they afford evidence of types not realized from intelligence or achievement tests which teachers should find valuable in the guidance and adjustment of their pupils.

*Achievement in the special subjects.* It is now possible to evaluate achievement and to diagnose disabilities with practical accuracy in the fields of whole numbers, fractions, decimals, percentage, mensuration, interest and business forms, and in problem-solving in arithmetic. The subject lends itself well to analysis and identification of specific skills, and thus to diagnosis. In other subjects, such as reading and language, a similarly exact identification of skills has not been accomplished, although some progress has been made in the



analysis of factors underlying achievement in broad skill areas in both reading and language. Tests capable of furnishing results accurate for individual diagnosis are now available for such reading skills as word meaning, sentence meaning, paragraph comprehension, rate of reading, and for certain of the more mechanical language skills, such as capitalization, punctuation, and language usage. Spelling and handwriting, two of the very important mechanical elements of written language, have both been analyzed and can be measured with reasonable success. The content subjects, such as the social sciences and the more exact sciences, because of vagueness in the statements of their aims and purposes, are extremely difficult to evaluate objectively. Then, too, mastery of the tool subjects, such as reading, language, and computational skills, enters into accomplishment in these fields to a very large degree.

Changes in educational emphasis from the vocational and practical to the cultural are creating an increased interest in measurement in the fine arts. The modern emphasis upon preventive measures in health education and adaptation of physical education to individual needs is motivation for many of the modern evaluative techniques of a non-test type in this area.

*General educational achievement.* While the emphasis throughout this volume is somewhat more on the measurement of the specific than the general aspects of school accomplishments, there is a recognizable need for the latter type of measurement. For general survey purposes, for evaluation of curricular content, and for later individual detailed diagnosis, such general achievement tests are valuable.

## Measurement and the total child

There are certain aspects of ability, accomplishment, skill, aptitude, character, and personality that unquestionably lend themselves to reasonably objective measurement. The emphasis placed on these measurable qualities frequently gives the impression that they represent the major elements in the total understanding of the child. Such is far from the case, however, for many of the intangibles of the child's personality are almost certainly of greater importance, although in many cases they are practically impossible to measure objectively. This merely means that the teacher must be made keenly aware of the fact that something lies beyond objective measurement.

He must see that appraisals of the child's total personality are basic to effective classroom teaching. He must recognize that many (probably most) of these vital appraisals must be made on the basis of keen observation and sympathetic analysis of his pupils. Even if the teacher were gifted with unusual observational and analytical power, superior native capacity, and natural sympathy, even if he were a four-year college or normal school graduate with graduate degrees in education, psychology, sociology, psychiatry, and medicine, he could not hope to comprehend more than a few problems of the child's personality. The important point here is that, while it is impossible to know all, it is not impossible for the teacher to be cognizant of and sensitive to these problems.

### **After testing, what?**

This question is in the back of the mind of every classroom teacher and every supervisor who has used standardized tests. Much of the early use of tests was futile, since such broad and vague phases of educational achievement were tested that, even though reliable results were obtained, nothing specific could be done about the situation. Furthermore, a great deal of the early use of tests in the classroom was a matter of satisfying curiosity. Teachers have a right to expect that something tangible will be given them in return for pupil time spent in testing. Pupils themselves may even have some rights in the matter. One way to insure this return is for the teachers themselves to take an active part in the program. A type of training, an attitude toward their profession, a clearer insight into the difficulties faced by their pupils, are thereby gained which may not come to them in any other way.

The results of supervisory tests given periodically for the purpose of checking the efficiency of pupil learning should be revealed to the teacher and the pupils in terms of specific suggestions for the further improvement of the situation. Instructional and diagnostic tests used by teachers in the classroom should furnish such specific information concerning the abilities and limitations of their pupils that a program of preventive and corrective instruction can be begun at once.



## 7 ORGANIZATION OF THIS BOOK

### Purpose of this book

The purpose of this book is twofold: (1) to interest the student of education in the possibilities of measurement and evaluation in education, and (2) to stimulate the teacher and supervisor to make more effective use of tests and other evaluative devices as integral parts of enlightened teaching practice. To accomplish this twofold purpose the reader is gradually introduced to the meaning and possibilities of measurement through the examination of some of the well-known current classroom practices. Chapter 2 briefly outlines certain historically important steps in the development of educational and mental tests. Tests are classified into their major types in Chapter 3, and a brief description is given of each type.

Chapter 4, which discusses at some length the characteristics or criteria of a good examination, is exceedingly important. It can most advantageously be studied after a reasonable understanding of certain statistical and correlational techniques has been established. A comprehensive understanding of the three most important criteria of a good examination depends upon the ability to interpret correlation coefficients. It is believed that a careful study of selected sections of Chapters 12, 13, and 14 will sufficiently acquaint the student with the meaning and uses of correlation for the immediate purposes of the discussion in Chapter 4.

Chapters 5 to 11 present the methods of constructing and the values and uses to the teacher of the major types of tests and evaluative techniques—standardized tests in Chapter 5, teacher-made essay and informal objective tests in Chapters 6 and 7, performance tests in Chapter 8, and evaluative techniques in Chapter 9. Intelligence and aptitude tests and personality instruments and techniques are discussed in Chapters 10 and 11.

Those especially interested in following to its logical conclusion the use of tests in the classroom will wish to study the remaining chapters with particular care, for here are presented the possibilities and the practical methods of using test results for analyzing and diagnosing the learning difficulties of pupils and the inauguration of preventive and remedial instruction in important school subjects.

## Study aids

The student who is genuinely interested in improving his understanding of many of the points presented in this volume will find much profit in the careful preparation of the discussion exercises at the end of each chapter. Those who are still more deeply interested in, and wish to pursue further, the problems of measurement in education will find the selected references at the close of each chapter of particular value. Because the field of educational measurements is so rich and the essential material is so extensive, it is impossible to compress into the pages allotted to this book even a good deal of the material that is considered by many to be fundamental.

Teachers, themselves expert in the technique of learning, know that passive reading, while yielding information and appreciation, does not develop easy, dependable skill in doing the thing described. To provide the opportunity for the student and the teacher actually to secure a more complete mastery of certain of these techniques, a *Work-Book in Educational Measurements and Evaluation* has been prepared as a companion volume for this treatment. In this *Work-Book* the reader solves practical problems of the type that the classroom teacher and supervisor face. Mastery of this text and a careful working of the projects in the *Work-Book* will practically guarantee to the reader an actual concrete experience with the major problems of a dynamic testing program calculated to be of the greatest service in the improvement of typical classroom situations.

## Topics for Discussion

1. What specific evidence is there that the idea of measurement in education is not entirely new but has been in the minds of teachers for many years?
2. What are some of the chief differences in the attitudes of teachers toward measurement and evaluation thirty years ago and today?
3. How far is the classroom teacher responsible for the understanding and use of educational tests?
4. Why is it a good thing for all educational tests to be subjected to sharp criticism by teachers?
5. Indicate several of the major characteristics of educational tests.
6. In what specific ways are informal objective tests and standardized tests alike and in what ways are they unlike?



7. Specify several things that educational tests, when properly used, do for the classroom teacher and his pupils.
8. Show how tests and examinations play an important part in the public relations contacts of the school.

### Selected References

- BROOM, M. E. *Educational Measurements in the Elementary School*. New York: McGraw-Hill Book Co., Inc., 1939. Chapters 1-2.
- DOOLITTLE, N. A. "Minimum Essentials of Measurement for the Classroom Teacher." *School and Society*, 69:403; June 4, 1949.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. Chapter 1.
- LEE, J. MURRAY. *A Guide to Measurement in Secondary Schools*. New York: D. Appleton-Century Co., Inc., 1936. Chapter 1.
- LINDQUIST, E. F., editor. *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapters 1-3.
- MCCALL, WILLIAM A. *Measurement*. New York: Macmillan Co., 1939. Chapter 1.
- MCCONN, MAX. "The Uses and Abuses of Examinations." *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Co., 1936. Chapter 9.
- MICHEELS, WILLIAM J., AND KARNES, M. RAY. *Measuring Educational Achievement*. New York: McGraw-Hill Book Co., Inc., 1950. Chapters 1, 3.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. Chapter 1.
- ORLEANS, JACOB S. *Measurement in Education*. New York: Thomas Nelson and Sons, 1937. Chapter 14.
- REMMERS, H. H., AND GAGE, N. L. *Educational Measurement and Evaluation*. New York: Harper and Brothers, 1943. Chapter 1.
- ROSS, C. C. *Measurement in Today's Schools*. Second edition. New York: Prentice-Hall, Inc., 1947. Chapter 1.
- TRAXLER, ARTHUR E., AND OTHERS. *Introduction to Testing and the Use of Test Results in Public Schools*. New York: Harper and Brothers, 1953. Chapters 1-2.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapter 1.
- WOOD, BEN D., AND HAEFNER, RALPH. *Measuring and Guiding Individual Growth*. New York: Silver Burdett Co., 1948. Chapters 1-7.

## ***Development of Educational and Mental Measurement***

THE DEVELOPMENT of educational and mental measurement from the time of the earliest historical records to the present is traced briefly in this chapter for the following areas and periods:

- A. Measurement to 1800.
- B. Educational testing from 1800 to 1900.
- C. Educational measurement and evaluation from 1900 to the present.
- D. Intelligence testing from 1800 to 1900.
- E. Intelligence measurement from 1900 to the present.
- F. Personality evaluation from 1800 to the present.
- G. Present status of educational and mental measurement and evaluation.

Measurement of human behavior with primary reference to the capacities and educational attainments of school children can well be divided roughly into three periods. During the first period, from the beginning of historical records down to about the nineteenth century A.D., educational measurements were naturally quite crude. Although the fact that individuals differ widely in their capacities and abilities has been recognized for several thousand years and educational measurement made formal entrance to the schools as early as medieval times, relatively little progress in educational testing was made until the nineteenth century. During the second period, embracing approximately the nineteenth century, educational meas-



urement began to assimilate from various sources the ideas and the scientific and statistical techniques which were later to result in the modern objective testing movement. The brief third period, dating from about 1900 to the present, has been characterized by tremendous advances in statistical techniques, in the measurement and evaluation of achievement, intelligence, and personality, and in the classroom use of test results.

## 1 MEASUREMENT TO 1800

### Early oral examinations

The first evidences of the oral examination are found in ancient literature. The story is told in the Old Testament (Judges 12:5-7) of the test the Gileadites gave to the enemy Ephraimites who wished to cross the Jordan. When asked to pronounce the word "Shibboleth," the Ephraimites could answer only with "Sibboleth," whereas people of the friendly tribes could respond with the correct pronunciation. Forty-two thousand Ephraimites were killed because they failed to pass this objective test.<sup>1</sup> Socrates, in a method he made famous, subjected his pupils to exhaustive and searching questioning. Oral quizzing, Socratic or otherwise, has undoubtedly been a part of classroom procedure from the beginnings of teaching activity—in fact, there have been and still are times when, for certain teachers, it constitutes practically the whole of the teaching act.

### Early written examinations

Written tests are probably of more recent origin than oral quizzes, but even they date back many centuries. As early as 2200 B.C., China had an elaborate national system of examinations for the purpose of selecting her public officials, and these examinations have been known down through the ages for their unusual severity. Confined in isolated cells for hours at a time, candidates were compelled to write lengthy papers or treatises on assigned topics.<sup>2</sup>

<sup>1</sup> Norma V. Scheidemann, "The Earliest Recorded Objective Test," *School and Society*, 20:702; June 1, 1929.

<sup>2</sup> W. A. P. Martin, *The Chinese: Their Education, Philosophy, and Letters*. Harper and Brothers, New York, 1881. p. 45-49.

## Recognition of individual differences

Individual differences among people have long been recognized. Plato, nearly four centuries B.C., divided his ideal society into the three classes of workers, protectors, and rulers. He believed that persons suited to each class should receive education for the fullest development of their personalities.<sup>3</sup> Quintillian, shortly after the start of the Christian era, wrote that masters should observe differences in ability and inclinations of persons they instruct, for the "forms of mind are not less varied than those of bodies."<sup>4</sup>

## Classification of personality types

Impressionistic methods of judging personality and of analyzing character have doubtless been in vogue for many centuries. They are based in the main on physiognomy, body build or glandular makeup, and divination. Representative are phrenology, astrology, palmistry, and graphology.<sup>5</sup>

## First educational tests

The first tests used for the measurement of the results or outcomes of education were probably not unlike certain of the performance tests of today, at least to the extent that they measured physical performance and that they were not paper-and-pencil tests.

Among various primitive tribes, in which the young men were taught to hunt, fish, and fight, the initiation ceremonies prerequisite to their admission to the ranks of adult males tested knowledge of tribal customs, endurance, bravery, and other skills and abilities thought necessary for tribal protection.<sup>6</sup>

The ancient Spartans, whose educational curricula for their youth stressed physical development and stoicism, conducted examinations as early as 500 B.C. in which the young men underwent painful or-

<sup>3</sup> Edgar W. Knight, *Twenty Centuries of Education*. Ginn and Co., Boston, 1940. p. 62.

<sup>4</sup> William Boyd, *The History of Western Education*. A. and C. Black, Ltd., London, 1921. p. 76.

<sup>5</sup> Henry E. Garrett, *Great Experiments in Psychology*, Third edition. Appleton-Century-Crofts, Inc., New York, 1951. p. 175-81.

<sup>6</sup> Charles Russell, *Standard Tests*. Ginn and Co., Boston, 1930. p. 14-15.



deals.<sup>7</sup> In ancient Athens, the stress upon athletics and aesthetic development led to evaluation by means of games and contests and of reading, writing, and singing ability.<sup>8</sup>

### First tests in the school

In medieval times, the oral examination was used in universities. The University of Bologna by A.D. 1219 and the University of Paris before the close of the thirteenth century required degree candidates to defend their theses orally. However, the written educational examination probably made its first appearance for educational use at Cambridge, England, in 1702.<sup>9</sup>

## 2 EDUCATIONAL TESTING FROM 1800 TO 1900

### Early educational tests in America

According to available records,<sup>10</sup> the first examinations of note in this country were those of Boston in 1845. Prior to that date the school committee had orally examined all Boston pupils, or at least those in the highest class in each school, annually. As the pupils increased in numbers, this task became onerous and eventually received only perfunctory attention. Finally, the sub-committee appointed to survey the grammar departments of the schools in 1845 decided to use written examinations in lieu of the time-consuming oral examinations. These subject examinations, in the fields of arithmetic, astronomy, geography, grammar, history, and natural philosophy, were used to rank the schools in order of merit.

This Boston examination project is truly a highlight in the history of education in the United States. It made a great impression on Horace Mann, who at that time was Secretary of the Massachusetts Board of Education.<sup>11</sup> As editor of the *Common School Journal*, he published extracts from the report and made many noteworthy com-

<sup>7</sup> *Ibid.* p. 16.

<sup>8</sup> Knight, *op. cit.* p. 52-53.

<sup>9</sup> Albert R. Lang, *Modern Methods in Written Examinations*. Houghton Mifflin Co., Boston, 1930. p. 2-3.

<sup>10</sup> Otis W. Caldwell and Stuart A. Courtis, *Then and Now in Education, 1845-1923*. World Book Co., Yonkers, N. Y., 1923. Chapters 1, 3.

<sup>11</sup> Since Horace Mann doubtless exerted considerable influence on the sub-committee, the examinations were probably reflections of his ideas.

ments<sup>12</sup> on the subject of examinations. He concluded that the new written examination was so superior to the old oral quiz that no school committee would ever lapse into the former inadequate and uncertain practice. The reasons advanced by Horace Mann in support of the written examination were as follows:<sup>13</sup>

1. It is impartial.
2. It is just to the pupils.
3. It is more thorough than older forms of examination.
4. It prevents the "officious interference" of the teacher.
5. It "determines, beyond appeal or gainsaying, whether the pupils have been faithfully and competently taught."
6. It takes away "all possibility of favoritism."
7. It makes the information obtained available to all.
8. It enables all to appraise the ease or difficulty of the questions.

Although these ideas were apparently those represented by modern tests, the instruments were inadequate. It is significant to note also that in successive issues of the *Common School Journal* Mann suggested most of the elements in examinations that are found in the modern measurement and evaluation movement.

### Early objective tests

To Rev. George Fisher, an English schoolmaster, goes the credit for devising and using what were probably the first objective measures of achievement. His "scale books," used in the Greenwich Hospital School as early as 1864, provided means for evaluating accomplishments in handwriting, spelling, mathematics, navigation, Scripture knowledge, grammar and composition, French, general history, drawing, and practical science. In such subjects as handwriting and drawing, where qualitative rather than quantitative evaluation was the custom, specimens of pupil work were compared with "standard specimens" to determine numerical ratings. The numerical values for spelling and other subjects to which quantitative measures of achievement were commonly applied depended upon errors in performance.<sup>14</sup>

<sup>12</sup> Horace Mann, "Boston Grammar and Writing Schools." *Common School Journal*, Vol. VII, No. 19; October 1, 1845. Also reported in: Caldwell and Courtis, *op. cit.* p. 237-72.

<sup>13</sup> Caldwell and Courtis, *op. cit.* p. 37.

<sup>14</sup> E. B. Chadwick, "Statistics of Educational Results." *The Museum, A Quarterly Magazine of Education, Literature, and Science*, 3:480-84; January 1864.



Although Fisher's "scale books" included the germ of many of the ideas that are incorporated in our present-day educational scales, his work produced no lasting results because he lived too far in advance of the thought and educational practice of his day.

### First objective tests in America

In America, the real inventor of the comparative test was Dr. J. M. Rice, who, in 1894,<sup>15</sup> hit upon the idea he developed so effectively that it became the foundation of objective measurement in education. Rice, having administered a list of spelling words to pupils in many school systems and analyzed the results, confounded the educators at the 1897 session of the Department of Superintendence of the National Education Association with the declaration that pupils who had studied spelling thirty minutes a day for eight years were not better spellers than children who had studied the subject fifteen minutes a day for eight years. Rice was attacked and reviled for this "heresy," and some educators even attacked the use of a measure of how well pupils could spell for evaluating the efficiency of spelling instruction. They contended that spelling was taught to develop the pupils' minds and not to teach them to spell. It was more than ten years later that Rice's pioneering resulted in significant attention to the objective method in educational testing.<sup>16</sup>

## 3 EDUCATIONAL MEASUREMENT AND EVALUATION FROM 1900 TO THE PRESENT

### First book on educational measurement

Thorndike brought out the first book dealing primarily with mental and educational measurements in 1904,<sup>17</sup> and both through this book and his later influence on his students became more than any other person responsible for the early development and popularization of standardized educational tests.

<sup>15</sup> Leonard P. Ayres, "History and Present Status of Educational Measurements." *The Measurement of Educational Products*. Seventeenth Yearbook of the National Society for the Study of Education, Part II. Public School Publishing Co., Bloomington, Ill., 1918. p. 11.

<sup>16</sup> *Ibid.* p. 12.

<sup>17</sup> Edward L. Thorndike, *An Introduction to the Theory of Mental and Social Measurements*. Teachers College, Columbia University, New York, 1904.

## First standardized achievement tests

Stone, a student of Thorndike's, published his arithmetic reasoning test, the first standardized instrument to make its appearance, in 1908.<sup>18</sup> Thorndike in 1909 published his *Scale for Handwriting of Children*—the first standardized achievement scale.<sup>19</sup> During the period 1909 to 1915, a series of arithmetic tests and five scales for measuring abilities in English composition, spelling, drawing, and handwriting were published.<sup>20</sup> It is interesting to note that only two of these pioneer instruments were tests, while the remaining five were scales.

Educators at first opposed the standardized test and derided the testers. However, the spread of standardized testing continued, under the stimulation of at least three early developments:

(1) The numerous important studies of the accuracy of school marks, revealing the fact that they are highly subjective and inaccurate, demonstrated the need for instruments that would yield more accurate measures of achievement.

(2) The surveys of certain of the larger school systems both stimulated the construction and use of tests and were influenced by the development of more objective devices for measuring the abilities of pupils.

(3) The development of educational measurements in research bureaus organized in many of the larger school systems, universities, and state departments of public instruction was influential in popularizing the use of educational tests. Although the pioneer and most of the early standardized tests were for use in the elementary school, it was not many years until the high school and even the college were well provided with such instruments.

## Development of informal objective examinations

The idea of the informal objective examination, referred to during its early days rather loosely as the "New-Type Test" and the "Ob-

<sup>18</sup> Cliff W. Stone, *Arithmetical Abilities and Some Factors Determining Them*. Contributions to Education, No. 19. Teachers College, Columbia University, New York, 1908.

<sup>19</sup> Edward L. Thorndike, "Handwriting." *Teachers College Record*, 11:83-175; March 1910.

<sup>20</sup> C. W. Odell, *Educational Measurements in High School*. Century Co., New York, 1930. p. 34-35.



jective Test," apparently was first publicly expressed by McCall,<sup>21</sup> whose article in 1920 first suggested that teachers do not need to depend solely upon standardized tests but that they can construct their own objective tests for classroom use. The pioneer book dealing almost entirely with this testing adaptation was published in 1924.<sup>22</sup> The informal objective test has since come into such wide use that a survey in 1936 of testing practices among 1600 high-school teachers widely distributed throughout the country showed that 74 per cent used the informal objective and an additional 10 per cent used a combination of the informal objective and essay examinations.<sup>23</sup>

### Later development of standardized achievement tests

The history of achievement measurement since the late twenties has been characterized mainly by increasing recognition of the fact that test results offer only one, although the major one, of the types of acceptable evidence on pupil achievement. This tendency toward evaluation, which is broader in scope than testing, has been accompanied by a strong trend toward more scientific use of measurement tools.

Although the contributions of Tyler have been significant in both the standardized testing and the informal objective testing movements, it is probably in the latter field that his influence was first felt. He outlined steps of procedure for test construction and validation which clearly pointed out the essential dependence of a program of achievement testing on the objectives of instruction and the recognition of forms of pupil behavior indicating attainment of the desired instructional outcomes.<sup>24</sup> Perhaps he more than any other single test specialist was responsible for the extension of achievement testing to the more intangible outcomes of instruction, for his contributions nearly twenty years ago doubtless did much to bring into being

<sup>21</sup> William A. McCall, "A New Kind of School Examination." *Journal of Educational Research*, 1:33-46; January 1920.

<sup>22</sup> G. M. Ruch, *The Improvement of the Written Examination*. Scott, Foresman and Co., Chicago, 1924.

<sup>23</sup> J. Murray Lee and David Segel, *Testing Practices of High-School Teachers*. U. S. Office of Education Bulletin, 1936, No. 9. U. S. Government Printing Office, Washington, D. C., 1936. p. 6.

<sup>24</sup> Ralph W. Tyler, "A Generalized Technique for Constructing Achievement Tests." *Educational Research Bulletin*, 8:199-208; April 15, 1931.

the broad modern conception of evaluation to replace the earlier and narrower concept of testing.<sup>25</sup>

### Development of evaluation instruments

The Eight-Year Study of member schools of the Progressive Education Association, completed some twelve years ago, affected measurement and evaluation practices markedly. The evaluation staff, working under Tyler's direction, developed a series of instruments for measuring such outcomes as logical reasoning, ability to apply principles in the sciences, ability to interpret data, and ability to interpret literature.<sup>26</sup> These and other instruments since made available, including those developed in the Cooperative Study of General Education, are designed to measure functional and relatively intangible outcomes in areas of behavior rather than more formal and tangible instructional outcomes in separate subjects or areas of the curriculum. A related trend is evidenced in the batteries of tests developed during the past ten or so years for the measurement of general educational development.

### Development of evaluative tools and techniques

Paralleling the development of paper-and-pencil tests has been the development of other evaluative tools and of techniques for measuring procedures involved in and products resulting from certain types of skill performances and various other aspects of behavior of the whole child. Prominent among such evaluative tools are the check list, the rating scale, the questionnaire, the pupil profile, the class record sheet, and the cumulative record. Evaluative techniques are represented by the anecdotal report, the interview, the case study, the sociogram, and observational analyses of group dynamics.

<sup>25</sup> Ralph W. Tyler, *Constructing Achievement Tests*. Ohio State University, Columbus, Ohio, 1934.

<sup>26</sup> Eugene R. Smith, Ralph W. Tyler, and others, *Appraising and Recording Student Progress*. Harper and Brothers, New York, 1942.



## 4 INTELLIGENCE TESTING FROM 1800 TO 1900

### Scientific recognition of individual differences

It was not, apparently, until 1796, that individual differences in mental abilities were first brought under not the microscope, but, literally, the telescope. It was in that year at the Greenwich Astronomical Observatory in England that one of the observers who recorded the instant of time at which stars crossed the lines on telescope lenses was discharged because his observations consistently differed slightly from those of his colleagues. In 1816, however, it was discovered by an astronomer who read an account of this incident that an error of observation, called the "personal equation," characterized the work of all observers and that the amount of error varied from person to person and also in the same person from time to time.<sup>27</sup>

### Scientific study of individual differences

Galton, with the publication of his *Hereditary Genius* in 1869, brought the scientific study of individual differences into focus, developed it further by instituting measurement of various human physical traits and motor abilities, and even investigated mental ability by methods which many years later became highly fruitful.<sup>28</sup>

### Foundations of statistical method

Galton's most important contribution to educational measurements was not in the field of individual differences, however, but in the derivation of statistical methods. Here, in devising a system of "standard scores" and in developing graphically the idea for an objective measure of relationship, the correlation coefficient, he furnished tools essential not alone to the development of educational and mental testing but also to scientific method in education. Pearson later formulated the method now most commonly used for calculating the correlation coefficient.<sup>29</sup>

<sup>27</sup> Anne Anastasi and John P. Foley, Jr., *Differential Psychology: Individual and Group Differences in Behavior*, Revised edition. Macmillan Co., New York, 1949. p. 7-8.

<sup>28</sup> Joseph Peterson, *Early Conceptions and Tests of Intelligence*. World Book Co., Yonkers, N. Y., 1925. p. 73-75.

<sup>29</sup> Garrett, *op. cit.* p. 269-72.

## Early attempts to measure intelligence

Dr. E. S. Chaille, an American physician, is credited as early as 1887 with the development of standards and simple tests for judging the mental levels of children to the age of three and with having implied, although not definitely used, the concept of mental age<sup>30</sup> as an index of mental maturity.

Cattell apparently first used the term "mental test"<sup>31</sup> in 1890, almost at the beginning of the period during which scientific method was first applied to the measurement of mental ability. Attempts during the last decade of the nineteenth century by Cattell and others to measure intelligence by means of physical characteristics, sensory acuity, and motor skills tests gave, for the most part, negative results.<sup>32</sup>

During the same period, Binet and his colleagues were experimenting in France with tests of a somewhat similar but less specific type. In 1895, Binet and Henri described ten types of tests which, differing from American tests mainly in the much greater complexity of behavior they would measure, they thought were likely to discriminate between levels of mental ability.<sup>33</sup>

## 5 INTELLIGENCE MEASUREMENT FROM 1900 TO THE PRESENT

### First individual intelligence tests

Binet and Simon brought out the first intelligence scale in 1905, devising it primarily for the purpose of selecting mentally retarded pupils who required special instruction. This pioneer individual intelligence scale utilized the basic idea of interpreting the relative intelligence of different children at any given chronological age by the number of tests of varied types and increasing levels of difficulty they could pass. These characteristics were all re-embodied in the 1908 and 1911 revisions of the *Binet-Simon Scale* and also are basic to most individual intelligence scales even today. The 1908 Revision

<sup>30</sup> Florence L. Goodenough, "An Early Intelligence Test." *Child Development*, 5:13-18; March 1934.

<sup>31</sup> J. McKeen Cattell, "Mental Tests and Measurements." *Mind*, 15:375-81; July 1890.

<sup>32</sup> Frank N. Freeman, *Mental Tests: Their History, Principles and Applications*, Revised edition. Houghton Mifflin Co., Boston, 1939. p. 58.

<sup>33</sup> Anastasi and Foley, *op. cit.* p. 15.



introduced the fundamentally important concept of mental age (MA) and provided means for obtaining it.<sup>34</sup>

### Individual intelligence tests in America

Goddard, Kuhlmann, and Terman all adapted the Binet-Simon tests to use with American children during the period from 1911 to 1916. Terman and his collaborators made the *Stanford Revision of the Binet Scale* available in 1916, and in 1937 followed it with a second and more complete revision. These revisions make use of the intelligence quotient (IQ), based on the relationship between a child's mental age and his chronological age.<sup>35</sup>

### Group intelligence tests

Although various psychologists had been working on a group intelligence test, and Otis was near the point of issuing such a test around 1917, the *Army Alpha* test, used for measuring and placing American army recruits and draftees during World War I, was the first group intelligence test to be published. The *Army Alpha* test, widely used for testing men who could read and understand English, was accompanied by *Army Beta*, a non-language test for use with illiterates and men who, although perhaps literate in a foreign language, could not read English.<sup>36</sup> These tests were widely used by educators after the close of the war.

Group intelligence tests began making their appearance almost immediately following the end of World War I, and the period from 1918 to the middle twenties was marked both by the publication of many such tests and by an upsurge of general interest in intelligence testing. Although the testing techniques have been refined considerably since then, the past quarter century has brought no outstanding changes in the methods of measuring general intelligence. The *Army General Classification Test* and the *Army Individual Test of Mental Ability* served functions in World War II closely similar to those of *Army Alpha* and *Army Beta* in World War I. Several other armed service branches developed comparable instruments for use in their programs of selection and classification.

<sup>34</sup> Freeman, *op. cit.* p. 86-88.

<sup>35</sup> *Ibid.* p. 101.

<sup>36</sup> *Ibid.* p. 113-35.

## Aptitude or specific intelligence tests

The measurement of aptitudes, or those potentialities for success in an area of performance that exist prior to direct acquaintance with that area, has been tied up with intelligence testing both fore and aft. Early attempts to measure general intelligence were by means of tests of many specific traits and aptitudes, but that approach was dropped when Binet showed that tests of more complex forms of behavior were superior. It was soon apparent, however, that general intelligence tests were not highly predictive of certain types of performance, especially in the trades and industries.

Münsterberg's aptitude tests for telephone girls and streetcar motormen in 1913 were followed by tests of mechanical aptitude, musical aptitude, art aptitude, clerical aptitude, and aptitude for various subjects of the high-school and college curricula prior to 1930.<sup>37</sup> Spearman's splitting of total mental ability into a general factor and many specific factors<sup>38</sup> had its influence on this movement, and accounted for the fact that aptitude tests are frequently called specific intelligence tests.

## Factored intelligence tests

With the development of factor analysis methods, largely within the past two decades, certain group factors of intelligence thought to differ from the specific factors or aptitudes and also from general intelligence have emerged.<sup>39</sup> These were first recognized in measurement practice by the introduction of separate linguistic and quantitative, or verbal and non-verbal, scores into certain tests of mental ability that continued to furnish a general measure of intelligence. In addition, several batteries for the measurement of primary mental abilities, differential aptitudes, and general aptitudes, each designed to distinguish several group factors of intelligence, have made their appearance during the last ten or twelve years.

<sup>37</sup> Goodwin Watson, "The Specific Techniques of Investigation: Testing Intelligence, Aptitudes, and Personality." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II. Public School Publishing Co., Bloomington, Ill., 1938. p. 365-66.

<sup>38</sup> Charles Spearman, "'General Intelligence' Objectively Determined and Measured." *American Journal of Psychology*, 15:201-93; April 1904.

<sup>39</sup> Godfrey H. Thomson, *The Factorial Analysis of Human Ability*. Houghton Mifflin Co., Boston, 1939. p. 14.



## 6 PERSONALITY EVALUATION FROM 1800 TO THE PRESENT

### Antecedents of modern personality tests

Personality testing had its antecedents in the work of Kraepelin and Sommer on free association tests during the last decade of the nineteenth century. Although free association tests have persisted to the present day, the questionnaire and rating scale methods used by Galton and others at still earlier dates became the dominant early methods of personality measurement in America.<sup>40</sup>

### Modern personality inventories

Woodworth devised a *Personal Data Sheet*, in reality an inventory of neurotic tendencies and emotional maladjustment, for use with American soldiers during World War I. This was probably the outstanding early contribution in this field.<sup>41</sup> A significant number of these structured personality inventories have been developed during the past thirty years for the measurement of adjustment, attitudes, and vocational interests.

### Projective methods

Jung in 1905 published a free association test designed to reveal emotional complexes.<sup>42</sup> Hartshorne, May, and their colleagues made exhaustive studies of conduct in largely unstructured or free response situations in the Character Education Inquiry.<sup>43</sup> Although the *Rorschach*, the first modern projective test, was introduced in 1921, it was not until some fifteen years ago that projective techniques employing such unstructured situations as inkblots and pictures came into wide use in the study of personality.<sup>44</sup> An outgrowth of psychiatry and academic psychology, these unstructured methods of

<sup>40</sup> Anastasi and Foley, *op. cit.* p. 22.

<sup>41</sup> Watson, *op. cit.* p. 368.

<sup>42</sup> *Ibid.* p. 368.

<sup>43</sup> Hugh Hartshorne, Mark A. May, and others, *Studies in the Nature of Character*, Volumes I-III. Macmillan Co., New York, 1928; 1929, 1930.

<sup>44</sup> Watson, *op. cit.* p. 369.

studying personality came to be termed projective methods only in 1939.<sup>45</sup>

## 7 PRESENT STATUS OF EDUCATIONAL AND MENTAL MEASUREMENT AND EVALUATION

Although measurement and evaluation, whether of achievement, intelligence, or personality, are still in a developmental stage, Monroe<sup>46</sup> stated that the movement, beyond its infancy in 1920, had reached early adulthood by 1945. He also commented that the fifty or more types of objective test items or item groups represented a marked extension and improvement in techniques of measurement and evaluation.

Reavis<sup>47</sup> reported Educational Records Bureau estimates that in 1944 approximately 60 million tests were administered to around 20 million persons in the United States. Many of these tests were used by the armed services and civil service, but a significant proportion were used in schools, colleges, and industry. Woodruff and Pritchard<sup>48</sup> indicated that in 1948 their test files included 1080 tests representing the output of 74 test publishers.

The measurement and evaluation aspects of the school program have markedly increased in scope and significance during the past score or so of years. Measurement and evaluation techniques now not only reflect developments in educational philosophy and psychology but also increasingly are furnishing evidence that aids school officials in charting the future course. Pupil guidance may be considered the central theme, for directly or indirectly all educational planning and procedures are designed to effect improvements in the education and in the guidance of the individual school child. The classroom teacher remains the key person in pupil measurement and evaluation. Measurement and evaluation specialists, subject-matter

<sup>45</sup> Helen Sargent, "Projective Methods: Their Origins, Theory, and Application in Personality Research." *Psychological Bulletin*, 42:257-93; May 1945.

<sup>46</sup> Walter S. Monroe, "Educational Measurement in 1920 and in 1945." *Journal of Educational Research*, 38:334-40; January 1945.

<sup>47</sup> William C. Reavis, "Testing Is Big Business." *School Review*, 55:259-60; May 1947.

<sup>48</sup> Asahel D. Woodruff and Maralyn W. Pritchard, "Some Trends in the Development of Psychological Tests." *Educational and Psychological Measurement*, 9:105-8; Spring 1949.



specialists, and specialists in areas of child behavior increasingly co-operate with and depend upon the classroom teacher in the development of new instruments, tools, and techniques for pupil appraisal.

### Topics for Discussion

1. What were some of the ancient forerunners of educational tests?
2. Show how educational testing had its origins centuries before standardized and informal objective tests were developed.
3. Discuss the early recognition of individual differences in mental ability and personality.
4. List and evaluate the most important ideas concerning examinations expressed by Horace Mann.
5. Discuss the "scale books" developed by Rev. George Fisher and compare them with modern educational scales.
6. What was the significance for objective measurement and for educational research of the contributions made by Dr. J. M. Rice?
7. What three important educational developments of the first two decades of the present century indirectly stimulated the growth of interest in educational measurements?
8. Who were the pioneers in the development of standardized educational tests? What was their influence on the measurement movement?
9. Indicate the part played by informal objective examinations in the development of educational measurement.
10. What influences contributed to the rise of evaluation instruments, tools, and techniques?
11. By what method did workers in the field of mental ability first seek to measure intelligence? How successful were their attempts?
12. Discuss the contributions of Binet and Simon to the intelligence-testing movement.
13. Briefly discuss the development of group intelligence testing from World War I to the present.
14. What types of abilities are measured by general intelligence tests? By specific intelligence or aptitude tests? By tests of group factors of intelligence?
15. Discuss the early attempts to measure personality objectively and the more recent structured personality inventories and projective methods.
16. Comment upon the status of educational and mental measurement and evaluation today.
17. Discuss the significance of measurement and evaluation for educational planning and practices.

## Selected References

- ANASTASI, ANNE, AND FOLEY, JOHN P., JR. *Differential Psychology: Individual and Group Differences in Behavior*. Revised edition. New York: Macmillan Co., 1949. Chapter 1.
- AYRES, LEONARD P. "History and Present Status of Educational Measurements." *The Measurement of Educational Products*. Seventeenth Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1918. Chapter 1.
- FREEMAN, FRANK N. *Mental Tests: Their History, Principles and Applications*. Revised edition. Boston: Houghton Mifflin Co., 1939. Chapters 1-8.
- GARRETT, HENRY E. *Great Experiments in Psychology*. Third edition. New York: Appleton-Century-Crofts, Inc., 1951. Chapters 11-13, 15.
- GERRERICH, J. RAYMOND, chairman. "Educational and Psychological Testing." *Review of Educational Research*, 20:1-97; February 1950.
- GOODENOUGH, FLORENCE L. *Mental Testing: Its History, Principles, and Applications*. New York: Rinehart and Co., Inc., 1949. Chapters 3-6.
- HUNT, THELMA. *Measurement in Psychology*. New York: Prentice-Hall, Inc., 1936. Chapter 3.
- MONROE, WALTER S. "Educational Measurement in 1920 and in 1945." *Journal of Educational Research*, 38:334-40; January 1945.
- PINTNER, RUDOLF. *Intelligence Testing*. New edition. New York: Henry Holt and Co., 1931. Chapters 1-3.
- ROSS, C. C. *Measurement in Today's Schools*. Second edition. New York: Prentice-Hall, Inc., 1947. Chapter 2.
- SARGENT, HELEN. "Projective Methods: Their Origin, Theory, and Applications in Personality Research." *Psychological Bulletin*, 42:257-93; May 1945.
- SCATES, DOUGLAS E. "Fifty Years of Objective Measurement and Research in Education." *Journal of Educational Research*, 41:241-64; December 1947.
- SEGEL, DAVID, AND GERBERICH, J. RAYMOND. "Overview of Educational and Psychological Testing, 1946 to 1949." *Review of Educational Research*, 20:5-16; February 1950.
- TYLER, RALPH W. "The Specific Techniques of Investigation: Examining and Testing Acquired Knowledge, Skill, and Ability." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1938. Chapter 29.
- WATSON, GOODWIN. "The Specific Techniques of Investigation: Testing Intelligence, Aptitudes, and Personality." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for



the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1938. Chapter 30.

WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapter 2.

WOODRUFF, ASAHEL D., AND PRITCHARD, MARALYN W. "Some Trends in the Development of Psychological Tests." *Educational and Psychological Measurement*, 9:105-8; Spring 1949.

WRIGHTSTONE, J. WAYNE. "Frontiers in Educational Research in the Measurement of Aptitudes and Achievement." *Journal of Educational Research*, 40:389-96; January 1947.

WRIGHTSTONE, J. WAYNE. "Trends in Evaluation." *Educational Leadership*, 8:91-95; November 1950.

## ***Educational and Mental Measuring Instruments and Techniques***

THIS CHAPTER presents a classification of the tests and other tools and techniques used in educational and mental measurement:

- A. General classification of educational and mental tests.
- B. Basic types of educational tests.
- C. Evaluative tools and techniques of an educational nature.
- D. General intelligence tests.
- E. Specific intelligence tests.
- F. Group-factor tests of intelligence.
- G. Performance tests of intelligence.
- H. Inventories and techniques for personality evaluation.

The measurement and evaluation of the whole child involve the use of many tests and other devices that cannot properly be called tests. Most of the remaining chapters of this volume are devoted to treatments of the various types of tests, non-test tools, and techniques characterized only briefly here.

### **1 GENERAL CLASSIFICATION OF TESTS**

#### **Educational and mental tests**

Modern tests are so varied in type and purpose that it is extremely difficult to classify them clearly. Tests can be classified in terms of their forms, their origins, their functions, and their content. In this



chapter tests are first classified broadly by function—educational, intelligence, and personality—and within major divisions are classified by whatever pattern seems most likely to familiarize the student with their major characteristics.

*Educational tests* have as their primary function the measurement of the results or effects of instruction and learning. On the other hand, *intelligence tests*, or psychological examinations, have as their purpose the measurement of pupil intelligence or mental ability in a large degree without reference to what the pupil has learned either in or out of school. *Personality tests* attempt to measure such intangible aspects of behavior as attitudes, interests, and emotional adjustment.

There is not complete uniformity of terminology with respect to educational tests and mental tests. Although the former have a commonly-accepted meaning, the latter are thought variously to include educational, intelligence, and personality tests, to include intelligence and personality but not educational tests, and even to mean the same thing as intelligence tests. Modern practice seems often to make use of the three-way classification—educational tests, intelligence tests, and personality inventories. As this distinction appears to be most satisfactory for the purposes of this book, it will be followed throughout this volume.

### Tests, scales, and scaled tests

Objective tests can be classified in a manner that cuts across the three fields of educational, mental, and personality testing—into tests and scales, and also scaled tests. This distinction is of some value, but at times it results in confusion since certain types of objective tests resemble scales or contain certain features of scales as an essential part of their construction.

In general terms, a *test* is an instrument designed for the measurement and evaluation of any knowledge, quality, or ability. It may measure degree or amount of achievement, mental abilities, or even such intangible qualities as personality and character traits. It may be made up of items of similar difficulty, items arranged in increasing order of difficulty, or items arranged in such other ways as by types of items or order of occurrence of topics in a course. Ordinarily the test is used in the classroom by the pupils.

A *scale* is a series of objective samples or products of different

difficulty or quality that have been arranged in a definite order or position, usually in ascending order of difficulty or merit. The samples are equally spaced on a scale of value, of difficulty, or of quality. Usually the scale is employed by the teacher as an aid in the evaluation of the particular product.

A *scaled test* combines certain properties of the test and the scale. If the items in a test are arranged in order of increasing difficulty, the instrument is a scaled test. The process of determining the difficulty of test items and arranging them in an ascending order on that characteristic is called *scaling*.

### Excerpts from Iowa Every-Pupil Test of Basic Skills in Language <sup>1</sup>

#### PART III. USAGE

**Directions:** In each of the following sentences there are two numbered words or phrases enclosed in brackets. If you think the *first* word or phrase is correct, place an **X** in the *first* box of the row that is numbered the same as the sentence; if you think the *second* answer is correct, place an **X** in the *second* box of the row.

Study the first two exercises carefully and see how they are marked on the answer sheet. Mark the other exercises in the same way.

1. The ball hit  $\left. \begin{array}{l} 1. \text{ I} \\ 2. \text{ me} \end{array} \right\}$ .
2. That  $\left. \begin{array}{l} 1. \text{ isn't} \\ 2. \text{ ain't} \end{array} \right\}$  the way to do it.
3. The glass is  $\left. \begin{array}{l} 1. \text{ broke} \\ 2. \text{ broken} \end{array} \right\}$ .
4. Why  $\left. \begin{array}{l} 1. \text{ don't} \\ 2. \text{ doesn't} \end{array} \right\}$  Polly come?
75. Either you or I  $\left. \begin{array}{l} 1. \text{ are} \\ 2. \text{ am} \end{array} \right\}$  going to win the prize.
76. There is the boy  $\left. \begin{array}{l} 1. \text{ who} \\ 2. \text{ whom} \end{array} \right\}$  I think will win the race.
77. Each of the men  $\left. \begin{array}{l} 1. \text{ was} \\ 2. \text{ were} \end{array} \right\}$  given two blankets.
78. If Jack  $\left. \begin{array}{l} 1. \text{ were} \\ 2. \text{ was} \end{array} \right\}$  as old as you, he would help you.

<sup>1</sup> H. F. Spitzer and others, *Iowa Every-Pupil Tests of Basic Skills, Test C, Basic Language Skills, Advanced, Form O*. Published by Houghton Mifflin Co., 1943.



The scaled test is illustrated by a sampling of items from the English usage section of the *Iowa Every-Pupil Tests of Basic Skills*. The items are the first four and the last four from the advanced level of the test and are designed for use in Grades 5 to 9.

### Speed and power tests

Tests of speed or rate and of power are used both in educational and intelligence testing, although personality inventories seem to involve neither the speed nor the power concept.

A *speed test* usually consists of items approximately equal in difficulty. Ordinarily such a test contains so many items that no pupil is able to finish in the working time allowed. Usually the items are so easy that there is no question about the pupil's ability to answer them correctly. The number of items answered correctly in the specified time is taken as a measure of the pupil's speed or rate of work. Thus, speed tests are measures of the speed and accuracy with which a pupil is able to respond to standardized items of a uniform degree of difficulty.

A *power test* consists of a series of items arranged in ascending order of difficulty, and hence is also a scaled test. It measures a pupil's ability to answer more and more difficult items within a given field. Usually no measure of the pupil's rate of work is secured, for the time allowed is sufficient for nearly all pupils to complete as many of the items as they are able to answer. In actual practice, however, the factors of power and speed are combined by taking as the pupil's score the number of items he answers correctly in the specified time. Theoretically, a pupil's score on a power test should represent the degree of difficulty of the most difficult item he is able to answer correctly, but such a score is so hard to obtain that the number of items answered correctly is generally taken as his score. A *work-limit test*, as the term is used in intelligence testing, is a power test on which the pupil may work until he is satisfied that he has done all he can or at least until practically every pupil in the test group has finished.

If a speed test may be compared to a race in which as many hurdles of uniform height as possible are to be cleared during a specified period of time, a power test may be compared to a contest in which the hurdles to be jumped regularly increase in height from very low at the start to such eventual height that no one is able to jump the

next hurdle. In the first case, the score (speed) would be expressed in terms of the number of hurdles jumped during the specified time. In the second case, the score (power) would be expressed as the height of the last and highest hurdle the individual was able to clear.

Many modern tests are really hybrids resulting from a combination of the power idea and the speed idea. They are made up of items arranged in ascending order of difficulty, but the resulting scores are expressed in terms of the number of items answered correctly in the specified working time. Since in achievement testing sufficient time is ordinarily allowed for at least 80 or 90 per cent of the pupils to finish, the speed factor does not receive much weight in the resulting scores. A *time-limit test*, however, common in the measurement of intelligence, is a power test given with such limited timing that no pupil is likely to complete it during the time allowed. Used in intelligence testing when a measure reflecting both power and speed is desired, the time-limit test has no direct counterpart in educational testing.

Furthermore, certain multiple-attribute tests of achievement combine the measurement of power and speed in one test or performance although not in one score. Thus, accuracy and speed in typewriting, legibility and speed in handwriting, and comprehension and speed in reading are dual, although not necessarily equally important, characteristics of a good performance. Inasmuch as stress on speed usually reduces quality and stress on quality typically reduces speed of performance, an optimum balance between the two characteristics is ordinarily desired. Even here, however, the relative demands for quality and speed may vary considerably both in school and in out-of-school situations where such performances are common, so that no fixed pattern of weighting for the power or quality and the speed scores is feasible.

### Verbal, non-verbal, and performance tests

Another classification cross-cutting educational, intelligence, and personality tests is that dependent on the degree to which words are used in test items and in pupil responses.

*Verbal tests*, by far the most common, are ordinarily of the pencil-and-paper variety although they may be oral or may even require identification of physical objects and materials presented. In any



event, words are used by the pupils in attaching meaning to, in responding to, or both in comprehending and in responding to, the test items. A test unless qualified or further described is ordinarily a verbal test of the pencil-and-paper type.

*Non-verbal tests*, again of the pencil-and-paper or even oral variety, are those in which pupils do not use words in attaching meaning to or in responding to test items. Tests involving the use solely of numbers, of graphical representations, or of three-dimensional objects and materials are of this type.

*Performance tests* are also non-verbal but they may require use of pencil and paper by the pupils in responding, they may require solely the manipulation of physical objects and materials, or they may require paper-and-pencil responses to physical objects and materials presented in certain ways. Such tests are commonly used with persons having serious language handicaps and in situations where certain types of skills are of greater importance than is verbalization ability.

Distinctions between verbal and language tests and non-verbal and non-language tests should perhaps be made here. A verbal test is necessarily a language test. But a test may involve language, either oral or written, in the instructions given to the pupils and nevertheless be non-verbal if the test itself and the pupil responses do not involve language. Some performance tests must be non-language as well as non-verbal, and hence involve the giving of instructions in pantomime.

## Teacher-made and standardized tests

A distinction most fundamental in educational testing, also applicable to personality measurement but not pertinent in intelligence testing, is that between the teacher-made and the standardized test. The teacher often constructs educational achievement tests and sometimes develops informal inventories or questionnaires for measuring such personality characteristics as interests and attitudes. Standardized tests occur in all three major areas of measurement—educational, intelligence, and personality.

The most common types of *teacher-made tests* are the oral, essay, and informal objective. The *oral test* is typically developed by the teacher in the classroom as the occasion warrants and consists of asking individual pupils questions to be answered orally. The *essay*

*test*, consisting of questions to which the pupils respond in writing, is also ordinarily used by a teacher in his own classes.

*Informal objective examinations*, most often prepared by a teacher for use in his own classes, may be constructed cooperatively by two or more teachers for use with their several classes in the same subject, or even by several persons for use throughout a large school system. Such tests may even be printed. They are, however, informal objective examinations unless procedures for standardization are carried out and the tests are made available for general use to interested persons outside of the school situation in which they originated. Illustrations of item types commonly used in informal objective tests are given later in this chapter.

A *standardized test* is composed of test or inventory items selected in the light of the particular type of achievement, mental ability, or personality trait the instrument is designed to measure. The items have necessarily been subjected to a preliminary tryout with a representative pupil group so that it became possible to arrange them in the desired manner with respect to difficulty and the degree to which they effected certain types of discriminations among groups of pupils. Such a test is accompanied by the appropriate type of table for transforming resulting scores into meaningful characterizations of pupil achievement, mental ability, or personality.

Both the informal objective test and the standardized test come under the broader heading of *objective tests*. When used in measuring educational achievement, the informal objective and standardized tests typically make use of the same item types. Both of these examination types are marked by three important features: (1) brevity of pupil response, (2) extensive sampling, and (3) absence of personal judgment in the scoring of the examinations. The pupil indicates his response by such simple physical reactions as underlining a word, encircling a number, filling an answer space, or writing a word or short phrase in an indicated place.

## Tests, non-test tools, and techniques

An important but perhaps fairly obvious distinction is that among tests, non-test instruments or tools, and techniques neither of a test nor of a tool nature. The distinction is important particularly in educational and personality measurement but also pertains in some degree to mental measurement.



*Tests*, discussed above, are used directly by the pupil. There are a number of *non-test tools*, such as the cumulative record, pupil profile, progress chart, class analysis chart, and report card, which are not used by the pupil who is being evaluated. In addition, there are a number of *techniques*, primarily observational in nature, such as the anecdote, the case study, the interview, sociometric methods, and techniques in the area of group dynamics.

## 2 EDUCATIONAL TESTS

Considered as educational here and throughout this book are all instruments and techniques designed to measure what the individual has learned both in and out of school. It is obviously impossible to be certain about the exact proportions of the attainments of a school pupil that are the result of direct classroom instruction, of the by-products of classroom and other school activities, and of the wide range of his out-of-school experiences. A rather wide variety of tests, of other instruments, and of techniques for measuring types of abilities not definitely taught in any classroom or even in the school should be considered educational, for the education of the child is not confined entirely to the hours he spends in school. This broad conception of educational measurement and evaluation underlies the point of view presented in this volume. Various aspects of educational testing are dealt with in greater detail in Chapters 5 to 9 and in the section of this book devoted to measurement and evaluation in various subject fields.

When examinations, other tools, and techniques are classified in terms of their form or structure, five types may be distinguished: (1) oral examinations, (2) essay examinations, (3) objective examinations and scales, (4) performance tests and scales, and (5) other evaluative instruments and techniques.

### Oral examinations

Oral questioning of pupil groups is used in the classroom for measuring recall of factual knowledges. Such questioning often constitutes a major part of the so-called recitation, in fact. It usually consists of asking pupils sequentially or in a somewhat random order to answer questions based on the assignment for the day and

of attempting to evaluate the quality of their responses. Oral examining may take many forms, some of which are inappropriate but others of which are sound as a measurement device. A section of Chapter 6 deals with the oral examination in some detail.

### Essay examinations

In the essay examination, a limited number of questions (usually five to ten) are stated by the teacher as a basis for written answers by the pupils. Typically, the questions are selected by the teacher to elicit essay-type responses on the subject matter of the course the individual pupils have learned. This type of examination frequently poses a question of the *who*, *when*, *where*, *what*, or *why* type, although it may ask pupils to name, to locate, to discuss, to evaluate, to distinguish between, to define or describe, to illustrate or explain, to give reasons for or causes of, or otherwise respond to more-or-less definite issues.

The essay-type examination is often used as a final examination or as a test over several weeks of course work. As such, it may be thought of as the essay examination proper. It is also often used in shorter form as a check on pupil preparation of assignments, in which situation it is usually known as a written quiz. This form of examination has both inappropriate and sound uses and may result either in accurate or inaccurate judgments concerning pupil achievement. A quite complete discussion of the essay examination and of means for insuring its accuracy as a measuring instrument when it is appropriately used appears in Chapter 6.

### Objective examinations and scales

As was pointed out in a preceding section of this chapter, the distinction between informal objective tests and standardized tests of achievement is concerned with matters other than the types of test items employed. The item forms used in the informal objective test are limited only by the degree of ingenuity employed by the teacher or teachers who construct it. Similarly, the quality of the informal objective test and its appropriate uses are below those of the standardized test only if the constructors lack the ability and the desire to construct a good test. The principles of objective test construc-



tion are so closely similar for these two types of tests that they are treated here under the more general heading of objective tests.

A tremendous variety of item types has been developed, and new adaptations are quite common. However, all objective items may be classified either as the *recognition* or the *recall* type. Recognition types, of which the alternate-response, multiple-choice, and matching forms are the most common, make only indirect demands upon the initiative of the pupil, inasmuch as the factual material basic to the issue in question is stated (or misstated) in the item. Recall types, however, of which the simple recall and completion forms are probably the most common, place demands upon the initiative and frequently the memory of the pupil by expecting him to supply and state the correct answer.

The illustrations below show how a factual knowledge can be measured by the three of the above types that are most brief in form. The first two are recognition and the third is recall in form.

1. The President of the United States in 1863 was Abraham Lincoln. Ⓓ F
2. The President of the United States in 1863 was (a) Ulysses S. Grant, (b) Millard Fillmore, (c) Abraham Lincoln, (d) Andrew Johnson, (e) Zachary Taylor. (c)
3. The President of the United States in 1863 was (Abraham Lincoln)

The tremendous variety of objective examination item types and the complexity of some of them makes impossible the presentation of more than a few of the most common forms here. A comprehensive treatment of this important type of examination is given in Chapters 5 and 7.

*Survey, inventory, and prognostic tests.* These three types of tests serve different purposes and are constructed on somewhat different lines, but all three may be considered general tests in the sense that their functions demand resulting scores which have general significance rather than highly specific or analytic meaning.

*Survey tests* are instruments that measure general achievement in certain subjects or fields of knowledge. They are used to test skills and abilities of widely varying types. Thus, a survey test might measure achievement in first-year algebra. Another, and broader, survey test might measure ability in all areas of mathematics at the high-school level. A still broader survey test might measure abilities in all of the major areas of the secondary-school course of study.

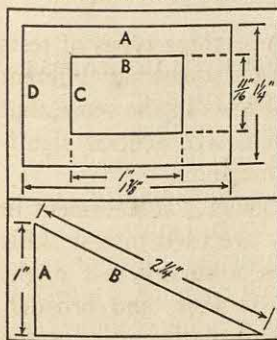
*Inventory tests* are similar to survey tests in specific school subjects, although the similarity is restricted to their form rather than to their use. Whereas survey tests are often used after instruction or during the instructional process, inventory tests are intended for use prior to instruction as an aid to the teacher in keying his instruction to the background learnings and levels of advancement of his pupils.

*Prognostic tests* are intended for use in the prognosis or prediction of future success in specific subjects of the school curriculum. As they usually test the background skills and abilities found to be prerequisite for success in the particular subject, prognostic tests are most common among subjects in which success can be rather well defined in terms of certain basic abilities. They also frequently test some of the aptitude factors that are not directly dependent upon previous training of a specific type. Therefore, prognostic tests, probably most closely related to aptitude tests but not unrelated to inventory tests, are more properly classified as educational tests than as intelligence tests, although they unquestionably do measure certain special aspects of intelligence. The accompanying illustration from the *Prognostic Test of Mechanical Abilities* shows how previous learning contributes to success on this type of test.

### Excerpt from Prognostic Test of Mechanical Abilities<sup>2</sup>

#### TEST II. READING SIMPLE DRAWINGS AND BLUEPRINTS

16-30. DIRECTIONS: The following are exercises in reading simple drawings and blueprints. Read each statement, look at the drawing, and write the letter that appears before the best answer on the line to the right of the statement. Do not use a ruler or scale.



16. The length of line A is:  
a  $1\frac{3}{8}$ " b  $1\frac{1}{2}$ " c  $1\frac{3}{4}$ " d  $1\frac{7}{8}$ " e 2" \_\_\_\_\_ 16
  17. The length of line C is:  
a  $\frac{1}{2}$ " b  $\frac{3}{4}$ " c  $1\frac{1}{4}$ " d  $1\frac{3}{16}$ " e  $1\frac{5}{16}$ " \_\_\_\_\_ 17
  18. The distance from line C to line D is:  
a  $\frac{1}{4}$ " b  $\frac{1}{2}$ " c  $\frac{7}{8}$ " d  $\frac{7}{16}$ " e 1" \_\_\_\_\_ 18
- 
19. The length of the longest line is:  
a 1" b  $1\frac{1}{2}$ " c 2" d  $2\frac{1}{8}$ " e  $2\frac{1}{4}$ " \_\_\_\_\_ 19
  20. The length of the shortest line is:  
a 1" b  $1\frac{1}{2}$ " c 2" d  $2\frac{1}{2}$ " e  $2\frac{3}{8}$ " \_\_\_\_\_ 20
  21. Line A is shorter than line B by:  
a  $\frac{1}{8}$ " b  $\frac{1}{4}$ " c  $\frac{1}{4}$ " d  $1\frac{1}{2}$ " e  $1\frac{5}{8}$ " \_\_\_\_\_ 21

<sup>2</sup> J. Wayne Wrightstone and Charles E. O'Toole, *Prognostic Test of Mechanical Abilities*, Form A. Published by California Test Bureau, 1946.



*Diagnostic and analytic tests.* Tests of diagnostic and analytic types are intended for the separate measurement of rather specific aspects of achievement in a single subject or field. Diagnostic tests measure somewhat narrower aspects of achievement than do analytic tests, so they may be thought of as serving specific and general diagnostic functions respectively.

*Diagnostic tests* yield measures of highly related abilities underlying achievement in a subject. They are designed to identify particular strengths and weaknesses on the part of the individual child, and within reasonable limits to reveal the underlying causes.

The relation of each skill to other skills and to the total process in the case of multiplication of fractions and mixed numbers is shown in the accompanying reproduction of Test VII of the *Compass Diagnostic Tests in Arithmetic*. The diagnostic procedure here is based on the assumption that mastery of the total process can be no stronger than the weakest link in the chain of related skills. Accordingly, each skill called into play in the total process so far as possible is isolated and measured. The parts of the test not reproduced here deal with reducing answers to best form, fundamentals of multiplication of fractions, and finding errors.

*Analytic tests* may be considered as general diagnostic tests. The term "diagnostic" as applied to educational tests has resulted in many misconceptions. Fundamentally, all tests may be considered diagnostic in the sense that they actually yield useful information about pupil achievement. However, the diagnosis afforded by many present-day tests is extremely general. Many so-called diagnostic tests are not diagnostic, but are merely analytic tests.

In contrast to the specific diagnosis that appears to be possible in the case of arithmetic is the general type of analysis that seems to climax the best efforts of test makers in the fields of language, reading, science, and the social studies. Attempts to analyze language and reading, for example, with a view to the construction of diagnostic instruments, immediately encounter the impossibility of relating to each other in any causal way the several phases of the subject on which achievement in it seems to depend. Causal relationships have not been established among such common factors in silent reading ability as word meaning, rate of reading, comprehension of facts, and ability to get the main idea. Consequently, in many subjects measures of different abilities are necessarily treated as independent and unrelated aspects of total ability.

Excerpt from Compass Diagnostic Tests in Arithmetic <sup>3</sup>

## PART 1—CHANGING MIXED NUMBERS TO IMPROPER FRACTIONS

Directions: Change the mixed numbers below to improper fractions.  
Study the samples before you begin to work.

SAMPLES:  $2\frac{1}{2} = \frac{11}{2}$      $8\frac{1}{4} = \frac{33}{4}$

$5\frac{3}{8} =$	$5\frac{1}{4} =$	$1\frac{1}{2} =$	$1\frac{1}{4} =$	$1\frac{2}{3} =$	$1\frac{7}{8} =$
$3\frac{3}{8} =$	$4\frac{3}{4} =$	$2\frac{3}{8} =$	$1\frac{4}{5} =$	$2\frac{3}{8} =$	$2\frac{1}{4} =$
$1\frac{1}{2} =$	$3\frac{1}{2} =$	$2\frac{1}{2} =$	$2\frac{2}{3} =$	$1\frac{1}{2} =$	$1\frac{1}{8} =$

Score on Part 1 = Number right = .....  
[Total possible score = 18 points]

## PART 2—CANCELLATION IN MULTIPLICATION OF FRACTIONS

Directions: Do all of the cancelling possible in the fractions below. Do not take time to finish the examples. Simply show all of the cancelling. Study the samples carefully before you begin to work.

SAMPLES:

$$\frac{1}{2} \times \frac{1}{2} = \frac{1}{4}$$

$$\frac{1}{1} \times \frac{2}{6} = \frac{1}{3}$$

$\frac{4}{5} \times \frac{1}{2}$	$\frac{6}{5} \times \frac{3}{10}$	$\frac{8}{5} \times \frac{5}{10}$	$\frac{1}{2} \times \frac{5}{6}$	$\frac{1}{1} \times \frac{1}{3}$	$\frac{1}{10} \times \frac{4}{5}$
$\frac{6}{5} \times \frac{5}{14}$	$\frac{5}{6} \times \frac{3}{4}$	$\frac{1}{2} \times \frac{4}{5}$	$\frac{1}{3} \times \frac{3}{4}$	$\frac{3}{10} \times \frac{5}{6}$	$\frac{4}{5} \times \frac{1}{4}$
$\frac{2}{3} \times \frac{4}{11}$	$\frac{3}{4} \times \frac{2}{5}$	$\frac{1}{15} \times \frac{5}{10}$	$\frac{3}{4} \times \frac{1}{2} \times \frac{1}{1}$	$\frac{3}{4} \times \frac{5}{6} \times \frac{1}{2}$	$\frac{1}{11} \times \frac{7}{5} \times \frac{3}{4}$

Score on Part 2 = Number right = .....  
[Total possible score = 18 points]

An illustration may serve to summarize the essential features of diagnostic and analytic tests. On certain points along the rim of the Grand Canyon there are lookout stations equipped with telescopes, each pointed and focused upon a specific spot of beauty or grandeur. From each of these a separate view of the beauty spots of the canyon is secured. From the composite of all of these views, a much more accurate appreciation of the total panorama is obtained, yet each view is quite independent of every other one. This is typical of the way in which tests of the analytic type operate. In distinct contrast with this example, the best way to illustrate the operation of the diagnostic type of test is to liken it to an inverted pyramid made of bricks. The removal or the crumbling of a single brick at any point in the wall will cause it to fall. The accompanying diagram may be helpful in clarifying the essential differences in these two types of tests.

<sup>3</sup> G. M. Ruch, F. B. Knight, H. A. Greene, and J. W. Studebaker, *Compass Diagnostic Tests in Arithmetic*, Test VII, Form A. Copyright by Scott, Foresman and Co., 1925. Reprinted by permission.



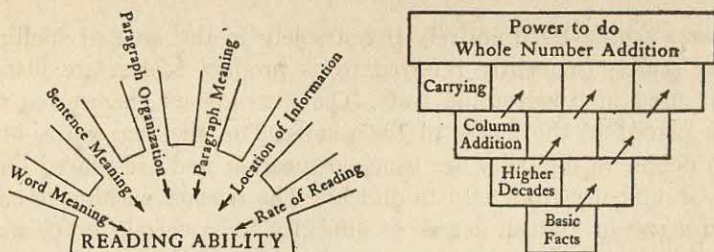


Fig. 1. Contrast between analysis and diagnosis

*Quizzes and mastery tests.* Instruments of these types are most often teacher-made, for they are ordinarily used in particular courses when the teacher sees need for them. They differ from most objective tests in length and in function more than in form. By nature the scope of these tests is quite restricted.

*Quizzes*, typically consisting of ten to fifteen true-false items or eight to ten simple recall items or problems, are used by many teachers on occasion during a portion of a class period for the primary purpose of determining whether or not the pupils have read assigned materials. Although they may be announced in advance, they frequently are given without forewarning to the pupils.

*Mastery tests* are designed to measure only those fundamental skills and abilities that all pupils supposedly have acquired, so the tests are at a very low level of difficulty for most pupils. Perfect scores are therefore commonly made by a majority of the pupils. These tests are typically constructed by the teacher for use in his own classes, although some of the workbooks designed for use in particular courses include tests of this type.

*Instructional and practice tests.* Tests of these types are sometimes constructed by the teacher, but they are also included in many of the workbooks provided particularly for elementary-school subjects and to some degree for high-school subjects. The scope of a test of this type is narrow, for typically only one aspect of a skill or one phase of a content area is covered. The exercises in such instruments may be either objective or semi-objective, for the primary value lies in their use as teaching aids.

Variously called instructional tests, practice tests, drill tests, and practice exercises, these instruments merit brief mention here because they are measuring instruments even though they are designed primarily for teaching rather than for testing purposes.

*Source scales.* Used entirely if not solely in the area of spelling, source scales, frequently referred to as product scales, are instruments used in constructing tests. The instruments themselves are never placed in the hands of the pupils. The spelling words of a given degree of difficulty are grouped together and are placed on a scale of difficulty from easy to difficult. The teacher wishing to construct a test of a given degree of difficulty for a certain grade may do so by selecting appropriate words from the source scale.

The following excerpt from the *Iowa Spelling Scales* for the eighth grade shows the words in Steps 12 and 15, for which the average percentages of misspellings are 58 and 34 respectively. On the average, the word "client" in Step 12 is misspelled by 62 per cent and the word "canvass" in Step 15 by 31 per cent of eighth-grade pupils.

#### Excerpts from Iowa Spelling Scales <sup>4</sup>

Step 12 — 58%	58%	55%
62%	all right	anticipating
client	alumni	circuit
convenient	anticipate	disappoint
council	assessment	equipped
immense	candidacy	immediately
permanent	continuous	
principle	fundamental	Step 15 — 34%
	geometry	37%
61%	girlie	definitely
accredited	physician	fraternally
characteristic	possess	
courteous	thorough	35%
losing		anniversary
scientific	57%	
	thoroughly	34%
60%		zephyr
analysis	56%	
correspondence	accompanying	33%
mortgage	acquaintance	chautauqua
Sabbath	auntie	X-ray
	originally	
59%	recommendation	31%
enthusiastic		canvass
lieutenant		
unusually		

<sup>4</sup> Ernest J. Ashbaugh, *The Iowa Spelling Scales*, Grade VIII. Published by Bureau of Educational Research and Service, State University of Iowa, 1944.



## Performance tests and scales

Tests of this type involve the manipulation or at least the use in some form of two-dimensional and perhaps primarily of three-dimensional objects. Paper and a pencil are sometimes necessary for the pupil in taking such a test, but in other situations the pupil actually constructs or manipulates some physical materials in the attempt to demonstrate his skill. Performance tests are dealt with in Chapter 8 of this volume.

*Object tests.* In tests of this type, physical objects are presented to the pupils for them to identify by name or type, for them to classify in some prescribed manner, or for them to use in identifying certain characteristics of the objects. Manipulation of the objects is not entailed unless handling is necessary in obtaining answers to the questions. Pupil responses are often made by the use of pencil and paper, but this situation differs in obvious ways from a paper-and-pencil test where the same objects are produced in photographs or line drawings. Object tests are most often used in testing ability to identify and discriminate among the various tools, materials, and specimens used in the industrial and practical arts and the physical sciences.

*Quality and rating scales.* In many school subjects, such as handwriting, composition, industrial arts, and the practical arts, pupils produce results in some tangible form through the application of their skills to assigned tasks. Their products differ in quality and in speed of production. The teacher's problem becomes one of evaluating the product and perhaps of recording the time taken in its production.

A *quality scale* is used in judging certain types of products in handwriting, lettering, artwork, shopwork, and other areas. Handwriting scales consist of a series of specimen performances exhibiting varying degrees of quality from the lowest to the highest. The specimens are arranged systematically in order of increasing quality. Usually the quality of each is described numerically. A quality scale of this type is used by matching the performance to be described with the specimen of the scale that most nearly resembles it in quality.

A *rating scale* or *score card* is used in evaluating other types of products, especially those in industrial arts and practical arts. Such a

scale and assigned numerical ratings thought to be appropriate for each characteristic of an excellent product is prepared in advance. The teacher then is able to rate numerically the quality of each pupil's production by observing, measuring, and otherwise judging his product on each rating-scale characteristic and by obtaining the summation for all characteristics.

## Check lists

A *check list* is most often used in evaluating the procedures used by a pupil in the performance of some assigned skill. Such check lists are useful in laboratory sciences and other performance areas in which certain tasks are most effectively performed by the use of sequences and techniques previously taught to the pupil. A check list of actions, both appropriate and inappropriate, is prepared in advance. The teacher observes each pupil separately as he attempts to perform the assigned task and keeps a running account of his procedures, both good and bad, in order of occurrence. The pupil's performance is then evaluated in terms of how closely it compares with the most efficient and direct procedures for reaching the goal he was seeking.

## Other evaluative instruments and techniques

To be treated here are the tests, non-test instruments, and techniques that are used in evaluating learning outcomes not closely related to specific school subjects and used in the presentation and summarization of other measurement results. Knowledge of their use constitutes at least as important a part of the teacher's equipment for measuring and evaluating pupils as is true of the more formal instruments discussed above. Evaluative instruments and techniques are the subject of Chapter 9 of this volume.

*Evaluative tests.* Although it is difficult and hazardous to attempt any distinction between evaluative tests and more formal tests, there are at least several types of tests which in form or in types of behavior measured probably should be considered evaluative in nature. In general, these evaluative tests are designed to measure some of the relatively intangible types of instructional outcomes.

Interpretive tests are represented by integrated units of test ma-



materials for the measurement of such relatively intangible outcomes as ability to interpret data, ability to interpret literature, ability to apply principles in the sciences, logical reasoning, and critical thinking. Tests of practices and activities, concerned with such out-of-school activities as health and safety practices and fiction reading, are also of this type. Certain tests of values in the areas of literature and reading, the fine arts, and recreations measure such intangible outcomes of school and out-of-school experiences as appreciations and satisfactions. In addition, some of the test batteries properly classified as a whole under survey tests have distinct evaluative features embodied in certain parts. Found in such batteries are basic skills tests in the portions not related directly to school subjects and the interpretive parts of tests of general educational development.

*Other evaluative tools.* Among these other tools used in pupil evaluation are the profile chart, the progress chart, the cumulative record, and the report card, all used with the individual pupil, and the class analysis chart, used both for an over-all evaluation of the group and for relating the status of the individual pupil to the over-all group picture.

*Evaluative techniques.* The interview and the questionnaire, both usually informal when used in pupil evaluation, are perhaps the most widely used techniques for evaluating educational outcomes. Both are, of course, used more formally for more specialized purposes. Some of the techniques discussed under personality evaluation later in this chapter, such as the anecdotal record and the case study, may also be used to measure educational outcomes even though their typical use is in the area of personality evaluation.

### 3 INTELLIGENCE TESTS

Intelligence tests measure what is perhaps most simply and most commonly described as ability to learn or ability to adapt oneself to new situations. Whereas achievement tests measure skills or abilities more or less directly, intelligence tests face the problem of measuring mental qualities indirectly in terms of the manner in which an individual's intelligence affects or conditions his behavior. It is sufficient here merely to comment upon this important distinction. Chapter 10 presents more fully the problems and techniques of intelligence testing.

## General intelligence tests

The most widely known tests of mental ability are usually referred to as general intelligence tests, although such other terms as general mental ability tests and psychological examinations have almost identical meanings. Other terms having similar meanings are general ability tests and aptitude examinations. General intelligence tests attempt to measure mental ability broadly enough, by the use of a wide variety of test situations in scaled order of difficulty, to obtain a measure representative of the individual's mental efficiency in general.

Results from general intelligence tests have so many uses, as in educational guidance, vocational guidance, sectioning of classes, and diagnosis, that it is impossible at this point to do more than mention this fact.

*Individual intelligence scales.* Intelligence tests that can be administered to only one person at a time are known as individual intelligence examinations. Such tests require the full attention of a trained examiner. Although the techniques for administering these tests are highly standardized, the examiner modifies the procedure in various ways according to the age, ability, and even sex of the pupil being tested. Since these instruments are usually in scaled form, and are frequently devised to cover a wide age range, they are often called age scales.

Individual intelligence tests only on occasional sections require the use of pencil and paper by the subject under examination. Some parts are even of a performance test nature. Many of the pupils' responses are given orally and are recorded by the examiner.

*Group intelligence tests.* Group tests of intelligence or general mental ability are usually paper-and-pencil tests that can be administered to a large group of persons at the same time. Group intelligence tests of the "omnibus" variety are ordinarily not divided into parts but have the items in mixed order with respect to the nature of the abilities they test and also sometimes with respect to their objective form. More commonly, however, group tests of intelligence have a number of different parts, each of which deals with a certain broad type of performance. In several of these tests two or more part scores are combined to obtain a verbal score and the remaining two or more part scores are combined to net a non-verbal score. When these two aspects of mental ability are measured by separate



tests, the instruments are ordinarily called verbal tests and non-verbal tests.

### Excerpts from Pintner General Ability Tests <sup>5</sup>

#### TEST 2. LOGICAL SELECTION

Pintner Verbal: Intermediate, 1 A

**Directions.** Look at the sample that follows.

**Sample.** A table always has — 1 flowers 2 tablecloth 3 legs 4 varnished top 5 ~~vine~~...

A table always has legs, which is number 3; so the third answer space is marked in the margin.

Read each statement. Find the thing it is most likely to have. Then mark the answer space in the margin which is numbered the same.

	1	2	3	4	5
1. A forest <i>always</i> has —	1 snow	2 trees	3 beasts	4 a forester	5 hunters
2. A sled —	1 boys	2 runners	3 ice	4 paint	5 wood
3. A horse —	1 tail	2 harness	3 shoes	4 stable	5 rider
4. A train —	1 windows	2 passengers	3 wheels	4 iron doors	5 diner
5. An orchestra —	1 hall	2 conductor	3 drum	4 instruments	5 audience
6. A game —	1 players	2 cards	3 tables	4 penalties	5 goals

The accompanying sample items from the *Pintner General Ability Tests, Verbal Series*, for the intermediate grades, illustrate one of the techniques used in group intelligence tests.

### Specific intelligence tests

In contrast to the general intelligence tests that attempt to measure broadly the ability to learn are the tests of specific intelligence that attempt to measure ability to learn in relatively narrow fields of subject matter or areas of performance.

**Aptitude tests.** Aptitude tests are frequently referred to as tests of specific intelligence. They attempt to measure the aptitude of a person, and often to forecast his probable future success, in certain school subjects or certain areas of performance. They are designed for use with persons who may or may not have had previous experience in the achievement areas with which they deal. Such tests attempt to measure the potentialities for success apart from those abilities resulting from specific training. Aptitude tests are not necessarily used for predictive purposes, although that is probably their most common use. They are found for such areas as English, foreign languages, music, art, mathematics, and the sciences,

<sup>5</sup> Rudolf Pintner, *Pintner General Ability Tests, Verbal Series, Intermediate, Form A*. Published by World Book Co., 1938.

and for such specific subjects as algebra, geometry, physics, and chemistry.

*Readiness tests.* Reading readiness tests have for some years been used with primary-school children in order to determine whether or not they have reached a level of maturity necessary for success in reading. Arithmetic readiness tests have been devised more recently for use in determining whether pupils have sufficient mental maturity to permit efficient learning of various arithmetic skills. Although there might be some question concerning the classification of readiness here, it seems that particularly for children entering school for the first time these tests more largely measure special mental abilities than the results of learning.

### Group-factor tests

Tests of group factors of intelligence have evolved since general intelligence and specific aptitude tests first made their appearance, and have recently attained considerable growth in usage. Group factors of intelligence are in a sense midway between specific and general intelligence.

*Bi-factor tests.* The two group factors most widely accepted and embodied in testing practices are the verbal and non-verbal factors referred to above. Essentially the same factors are called linguistic and quantitative in one psychological examination, whereas in still another mental capacity test they are termed language and non-language.

*Multi-factor tests.* Still newer in testing practice are the several tests now available for measuring from six to eleven, and in one test considerably more, group factors of intelligence. Perhaps the term most often used to characterize these multiple group factors of intelligence is primary mental abilities. Representative of such factors are spatial, numerical, manual dexterity, memory span, induction, and deduction.

### Performance tests

Performance tests are of several types, which cut across the classifications of intelligence tests given above. Some are individual and others are group tests. Some measure general intelligence and others measure specific aptitudes. The term usually designates tests for



which motor or manual responses rather than verbal or written responses are required of the pupil. These tests are often devised for use with illiterates, backward children, persons who are unfamiliar with English although they may read and speak a foreign language, and handicapped persons of various types. They frequently involve pantomime rather than verbal or printed directions and are usually at a rather low level of difficulty.

One type is a group paper-and-pencil test of general intelligence requiring no handwriting, given to illiterates and others not able to read and speak English with ease. Another type of general intelligence test given to one person at a time requires such performances as fitting of blocks into form boards, putting together of what resembles a jigsaw puzzle, and imitating actions of the examiner. Still others, making use of manipulative tests similar to the above except that they require more dexterity and place a premium upon speed of response, are individual tests used in the measurement of certain types of mechanical aptitude for adolescents and even adults.

#### 4 PERSONALITY INVENTORIES AND EVALUATIONS

Although psychologists are in agreement that the common conception of personality is not psychologically sound, they are not in agreement concerning the real meaning of the term. They do, however, believe that personality has to do with the total behavior of the individual, both that which can and that which cannot be observed. In the discussion that follows, four of the types of behavior generally classified under personality which seem to be most useful concepts to the teacher are discussed. These, as well as some other types of behavior usually listed under personality, are discussed more completely in Chapter 11.

#### Attitudes scales

The attention that has recently been called to attitudes by several nationwide surveys of public opinion illustrates the educational importance of attitudes. Attitudes are formed, crystallized, and sometimes modified or changed in the home, the church, on the playground, and elsewhere, as well as in the school.

Attitudes scales are of several types, but they frequently are based on a two-, three-, or five-point scale of agreement-disagreement with

statements concerning controversial issues or at least issues on which opinions may readily differ. Some such scales deal with a specific issue, such as attitude toward the Chinese. Others are generalized, and may deal equally well with attitudes toward any racial group, or some other general quality.

The results from the measurement of attitudes are useful in a variety of ways both in school and in social situations, for certainly attitude changes occur as one type of instructional outcome and the attitudes of pupils undoubtedly influence their adjustment in the school.

### Interests inventories

The interests of different individuals vary tremendously. Not only are the individual's fields of interest sometimes obscured intentionally or unintentionally by his behavior, but in some instances his real interests may be unknown to him. Interests questionnaires use techniques somewhat similar to those for attitudes testing, and frequently request indications of the presence of interest or the degree of interest a person has in various occupations, modes of behavior, types of activity, kinds of reading, and types of recreation, to name only a few. Results from interests inventories are rather widely used in vocational guidance, and also have uses for the teacher in aiding him to adapt his instruction to pupil interests.

The accompanying illustration from one of the interest parts of the Pressey *Interest-Attitude Test* shows one method of measuring interests in a variety of things.

### Adjustment inventories

Adjustment inventories attempt to measure emotional adjustment primarily. Known by a variety of names—personality tests, personality inventories, personality schedules, adjustment inventories, and in various other ways—they ask the pupil to respond objectively to items probing his behavior, his likes and dislikes, his environment, and many other aspects of his life. A major purpose of such instruments is to locate those abnormalities and peculiarities of behavior, neurotic tendencies, and various other types of maladjustment that should receive immediate attention if the individual possessing them is to become a well-adjusted adult.



Excerpts from Pressey Interest-Attitude Test, Test III <sup>6</sup>

**Directions:** Below is a list of things that people often like or are interested in. Place a cross (X) on the dotted line in front of everything which YOU like or in which YOU are interested. Place two crosses (XX) in front of everything in which you are VERY MUCH interested . . . which you like VERY MUCH. You may mark as many or as few words as you wish. But be sure to mark everything which you like or in which you are interested.

- |                   |                     |                       |
|-------------------|---------------------|-----------------------|
| 1.....artist      | 31.....card parties | 61.....dress          |
| 2.....drawing     | 32.....dancing      | 62.....reading        |
| 3.....cartoonist  | 33.....doctors      | 63.....children       |
| 4.....movie star  | 34.....fashions     | 64.....professors     |
| 5.....engineers   | 35.....leaders      | 65.....science        |
| 6.....comedies    | 36.....photography  | 66.....studying       |
| 7.....riding      | 37.....poker        | 67.....social affairs |
| .....horseback    | 38.....society      | 68.....coffee         |
| 8.....soldiers    | 39.....university   | 69.....cards          |
| 9.....typewriting | 40.....auto driving | 70.....waltzes        |
| 10.....carnival   |                     |                       |

The accompanying excerpt from the Student Form of the Bell *Adjustment Inventory* illustrates items dealing with the (a) home, (b) health, (c) social, and (d) emotional adjustment of the individual.

Excerpts from Bell Adjustment Inventory, Student Form <sup>7</sup>

## DIRECTIONS

Are you interested in knowing more about your own personality? If you will answer honestly and thoughtfully all of the questions on the pages that follow, it will be possible for you to obtain a better understanding of yourself.

There are *no right or wrong* answers. Indicate your answer to each question by drawing a circle around the "Yes," the "No," or the "?". Use the question mark only when you are certain that you cannot answer "Yes" or "No." There is no time limit, but work rapidly.

If you have *not* been living with your parents, answer certain of the questions with regard to the people with whom you have been living.

- |     |     |    |   |   |
|-----|-----|----|---|---|
| 1a  | Yes | No | ? | Do you day-dream frequently?  |
| 2a  | Yes | No | ? | Do you take cold rather easily from other people?                       |
| 3a  | Yes | No | ? | Do you enjoy social gatherings just to be with people?                  |
| 4a  | Yes | No | ? | Does it frighten you when you have to see a doctor about some illness?  |
| 5a  | Yes | No | ? | At a reception or tea do you seek to meet the important person present? |
| 6a  | Yes | No | ? | Are your eyes very sensitive to light?                                  |
| 7a  | Yes | No | ? | Did you ever have a strong desire to run away from home?                |
| 8a  | Yes | No | ? | Do you take responsibility for introducing people at a party?           |
| 9a  | Yes | No | ? | Do you sometimes feel that your parents are disappointed in you?        |
| 10a | Yes | No | ? | Do you frequently have spells of the "blues"?                           |

<sup>6</sup> S. L. Pressey, *Interest-Attitude Test*. Published by Psychological Corporation, 1933.

<sup>7</sup> Hugh M. Bell, *The Adjustment Inventory*, Student Form. Published by Stanford University Press, 1934.

## Evaluative techniques

Paper-and-pencil instruments are useful in measuring attitudes, interests, and even adjustment, but they cannot be used in measuring conduct in the many life and school situations where paper and a pencil are not natural parts of the environment. A number of these techniques, most of them observational, are used in individual pupil evaluation and several are intended for use in the evaluation of group behavior.

*Evaluation of individual behavior.* Several techniques have been evolved for noting the overt behavior of the whole child in natural as well as controlled situations. Among such techniques of greatest concern for the teacher are anecdotal records, a wide variety of projective methods, and the case study. The anecdotal record consists of a factual narrative of a particular situation observed by the recorder in which the pupil about whom the anecdote is written played a significant and relatively unique role. In the projective method, the child's behavior is observed and interpreted by a trained psychologist in a controlled situation where the child must react in some observable manner to materials presented by the psychologist. The case study, constituting a summary of a variety of observations and measurements of the individual, involves the assembly, integration, and interpretation of all important and obtainable information concerning the origins, background, environment, and status of the pupil.

*Evaluation of group dynamics.* Included in the evaluation area of group dynamics are two techniques of significance for the teacher. The first, the sociogram, employs paper-and-pencil sociometric methods for determining group patterns of behavior and the place of the individual within the group. Analyses of group interactions by direct observation in controlled situations again permits evaluation both of group behavior and of individual conduct within the framework of the social situation.

## Topics for Discussion

1. Distinguish the three general types of tests—educational, mental, and personality.
2. Distinguish between tests and scales. What are scaled tests?
3. Indicate the major differences between speed and power tests.
4. Distinguish among verbal, non-verbal, and performance tests.



5. Briefly characterize the four forms of educational tests—oral examination, essay examination, objective examination, and performance test.
6. Indicate the major characteristics of survey and prognostic tests. Of diagnostic and analytic tests.
7. Illustrate several types of items used in objective examinations.
8. Discuss several types of educational performance tests.
9. For what achievement areas are source scales and quality scales provided? How are these scales related to tests?
10. What types of evaluative tests and techniques are used in educational measurement?
11. Distinguish among tests of general intelligence, of specific intelligence, and of group factors of intelligence.
12. Briefly note some of the differences between individual and group tests of general intelligence.
13. What do aptitude and readiness tests measure? For what fields are they provided?
14. What do group-factor tests measure? In what areas are they provided?
15. Briefly indicate the major characteristics and uses of performance tests of mental ability.
16. Briefly characterize attitudes scales, interests inventories, adjustment inventories, and other evaluative techniques used in personality measurement.

## Selected References

- BELL, JOHN E. *Projective Techniques: A Dynamic Approach to the Study of Personality*. New York: Longmans, Green and Co., 1948. Chapters 1, 25.
- BINGHAM, WALTER V. *Aptitudes and Aptitude Testing*. New York: Harper and Brothers, 1937. Chapters 4-5.
- CATTELL, RAYMOND B. *Description and Measurement of Personality*. Yonkers, N. Y.: World Book Co., 1946. Chapter 1.
- COOK, WALTER W. "Tests—Achievement." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 1461-78.
- CRONBACH, LEE J. *Essentials of Psychological Testing*. New York: Harper and Brothers, 1949. Chapters 11-12.
- ENGELHART, MAX D. "Examinations." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 407-14.
- FREEMAN, FRANK N. *Mental Tests: Their History, Principles and Ap-*

- plications*. Revised edition. Boston: Houghton Mifflin Co., 1939. Chapters 1-8.
- FRYER, DOUGLAS. *The Measurement of Interests*. New York: Henry Holt and Co., 1931. Chapter 10.
- GOOD, CARTER V., editor. *Dictionary of Education*. New York: McGraw-Hill Book Co., Inc., 1945.
- GREENE, EDWARD B. *Measurements of Human Behavior*. Revised edition. New York: Odyssey Press, 1952. Chapter 2.
- HULL, CLARK L. *Aptitude Testing*. Yonkers, N. Y.: World Book Co., 1928. Chapter 3.
- HUMPHREYS, LLOYD G., AND BOYNTON, PAUL L. "Intelligence and Intelligence Tests." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 600-12.
- HUNT, THELMA. *Measurement in Psychology*. New York: Prentice-Hall, Inc., 1936. Chapter 2.
- LEE, J. MURRAY. *A Guide to Measurement in Secondary Schools*. New York: D. Appleton-Century Co., Inc., 1936. Chapter 2.
- LINCOLN, EDWARD A., AND WORKMAN, LINWOOD L. *Testing and the Use of Test Results*. New York: Macmillan Co., 1935. Chapter 2.
- MICHEELS, WILLIAM J., AND KARNES, M. RAY. *Measuring Educational Achievement*. New York: McGraw-Hill Book Co., Inc., 1950. Chapter 2.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. Chapter 2.
- OLSON, WILLARD C. "Personality." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 806-17.
- ORLEANS, JACOB S. *Measurement in Education*. New York: Thomas Nelson and Sons, 1937. Chapters 4-5.
- PINTNER, RUDOLF. *Intelligence Testing*. New edition. New York: Henry Holt and Co., 1931. Chapters 6-7.
- QUILLEN, I. JAMES, AND HANNA, LAVONE A. *Education for Social Competence: Curriculum and Instruction in Secondary-School Social Studies*. Chicago: Scott, Foresman and Co., 1948. Chapter 13.
- REMMERS, H. H., AND SILANCE, E. B. "Generalized Attitude Scales." *Journal of Social Psychology*. 5:298-312; August 1934.
- STAGNER, ROSS. "Attitudes." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 77-84.
- SYMONDS, PERCIVAL M. *Diagnosing Personality and Conduct*. New York: D. Appleton-Century Co., Inc., 1931. Chapters 5-9.
- TERMAN, LEWIS M., AND MERRILL, MAUD A. *Measuring Intelligence: A Guide to the Administration of the New Revised Stanford-Binet Tests of Intelligence*. Boston: Houghton Mifflin Co., 1937.
- THURSTONE, L. L., *Primary Mental Abilities*. Chicago: University of Chicago Press, 1938.



- THURSTONE, L. L., AND CHAVE, E. J. *The Measurement of Attitude*. Chicago: University of Chicago Press, 1929.
- THUT, I. N., AND GERBERICH, J. RAYMOND. *Foundations of Method for Secondary Schools*. New York: McGraw-Hill Book Co., Inc., 1949. p. 162-90, 238-66, 323-47, 380-98.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapter 3.
- WEITZMAN, ELLIS, AND McNAMARA, WALTER J. *Constructing Classroom Examinations*. Chicago: Science Research Associates, 1949. Chapters 3-4.
- WRIGHTSTONE, J. WAYNE. "Evaluation." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 403-7.

## *Essential Qualities of a Good Measuring Instrument or Technique*

THE FOLLOWING aspects of the criteria or distinguishing characteristics of a good examination, non-test instrument, or evaluative technique are discussed in this chapter:

- A. Validity as an essential characteristic of good measurement or evaluation.
- B. Curricular and statistical validity.
- C. Reliability as an aspect of validity.
- D. Methods of determining and estimating reliability.
- E. Dependence of reliability upon adequacy and objectivity.
- F. Administrability, scorability, and economy as practical criteria.
- G. Comparability in the use of test and evaluation results.
- H. Utility as an over-all criterion of good measurement.

The selection of any standardized educational test, mental test, or personality inventory requires careful consideration of the characteristics of a good examination. Similarly, the construction of any test or non-test instrument and the preparation of evaluative techniques, whether educational, mental, or in the area of personality, require careful consideration of the characteristics of good measurement. Although the characteristics of a good examination, tool, or technique can be listed and classified in many different ways, test specialists are in general agreement concerning the aspects that should receive attention in selecting or constructing them. The cri-



teria discussed below undoubtedly represent the most important considerations to be taken into account.<sup>1</sup>

It is recommended that the student refer frequently to the discussion of Chapter 14 on the statistical methods of determining test validity and reliability in connection with the study of these two exceedingly important criteria. An adequate understanding of these criteria depends on both their theoretical and their statistical aspects.

## 1 VALIDITY

Validity is the most important characteristic of a good examination, for unless a test is valid it serves no useful function. *The validity of an examination depends on the efficiency with which it measures what it attempts to measure.* A test must, therefore, accomplish the purpose the user has in mind in order to satisfy this fundamental criterion for all testing. In fact, the uncritical acceptance of an invalid test by a teacher for performing a desired function might easily result in serious injustice to the pupils. Accordingly, teachers cannot be too careful in assuring themselves of the validity of the tests they use. For example, a teacher who used a test that measures only knowledge of facts in a course in American history would not be correct in drawing conclusions, on the basis of the results, about the abilities of his pupils to apply historical facts to the reasoned interpretation of events.

It follows, also, that a test must be used with pupils who possess the proper intellectual maturity and background of experience for taking the test if it is to possess validity. For example, a standardized arithmetic survey test intended for use with pupils in Grades 6 to 9 might be invalid for use with most of the pupils in Grade 5 and probably with all pupils in the lower grades.

Lindquist<sup>2</sup> illustrated validity by pointing out that a test of high validity for ranking high-school pupils on general achievement in

<sup>1</sup> Although the discussion of this chapter is typically in terms of test or examination criteria, in order to avoid cumbersome wording, the reader should bear in mind the fact that the broad and general rather than the narrower, special use of the term is intended. The criteria should be interpreted as applying in only slightly modified form to non-test evaluative tools and to evaluative techniques as well as to tests and examinations.

<sup>2</sup> Herbert E. Hawkes, E. F. Lindquist, and C. R. Mann, editors, *The Construction and Use of Achievement Examinations*. Houghton Mifflin Co., Boston, 1936. p. 21-22.

United States history would have constantly decreasing validities for testing college students over a course in the same subject, for testing high-school pupils over a course in economic history of the United States, for predicting future success in a secondary-school English history course, for diagnosing weaknesses in abilities in United States history, for measuring general intelligence, and, finally, as a basis for assigning course marks in manual training.

Validity is, therefore, a specific rather than a general criterion of a good examination. It is specific in the sense that a test may be highly valid for use in one situation and highly invalid for use in another manner. It is specific, also, in the sense that a test may be valid for use with one group of pupils but not for use with a different pupil group. *Tests cannot correctly be described as valid in general terms, but only in connection with their intended use and at the intended ability level of the pupils.*

There is a difference in the concept of validity which should be applied in the consideration of standardized and informal objective examinations. It is readily apparent that the teacher better than anyone else knows the content and emphases of the course he has taught. Therefore, in that sense, he is the person best qualified to construct a valid test for his course. However, it is frequently true that the makers of standardized tests are better able than many, if not most, classroom teachers to determine what commonly are, and perhaps what should be, the content and emphases in courses for which they construct and standardize tests. Therefore, it seems reasonable to conclude that insofar as test content is concerned the teacher is the person best qualified to test the attainment of the desired outcomes in his particular class, but that the standardized test affords a superior means for determining how well his pupils have attained the core outcomes that are most widely recognized as being desirable in the particular course he is teaching. This difference in the application of the concept of validity is the result of the fact that no two teachers teach *exactly* the same course and that no one teacher teaches *exactly* the same course twice during his lifetime. Although this is particularly true for courses in the contemporary social studies and literature, and in the sciences, in which new content must be introduced constantly to keep abreast of developments, it is true even of such subjects as mathematics, in which the methods used and classroom problems that arise may well differ from semester to semester even though the basic content may be largely unchanged.



Three types of test validity are discussed here: (1) curricular validity, (2) statistical validity, and (3) psychological and logical validity. Of these three, the first is by far the most important, for in the final analysis any method of test validation must be based on relatively subjective judgments concerning the degree to which an examination covers the proper ground. Statistical validity, in turn, is a more widely used and probably a more important concept than psychological and logical validity.

### Curricular validity

The first of the three types of methods used primarily in determining the validity of educational tests is curricular validation. A teacher who carefully and thoughtfully selects a standardized test or constructs an informal objective examination or any other evaluation instrument for his class is attempting to insure curricular validity by making certain that the test deals with the types of educational outcomes he wishes to measure and is at the proper level of difficulty for his pupils. There are various sources of evidence to guide the teacher in considering test validity from the curricular standpoint. Among these are textbooks, courses of study, reports of national committees, and the writings of subject and test specialists. The idea in each case is that analysis of these source materials will furnish evidence concerning the thinking of qualified educators on questions dealing with course objectives and that such an analysis affords an important objective basis for determining the outcomes to test.

*Textbook and course of study analyses.* The major weakness in the analysis of textbook and course of study content as a validation method is that it tends to perpetuate faulty and inadequate curricular content if such defects exist in the source materials. It does not look beyond present practices. On the other hand, the overlapping of instructional material that is common to a large number of textbooks and courses of study almost certainly represents important content.

*Recommendations of committees and subject and test specialists.* Reports of national committees and writings of subject and test specialists often serve as good guides to content in educational test construction. Such reports and recommendations are usually based on carefully formulated statements of instructional objectives and of

desired learning outcomes stated in terms of pupil behavior. These source materials often provide excellent foundations for standardized and informal objective test construction, and the use of modern reports and writings by recognized committees and specialists is unlikely to result in perpetuating the errors in past practices.

*Local determination of objectives and outcomes.* Much emphasis in modern schools is properly placed on teacher participation in the formulation of instructional objectives and of resulting behavioral outcomes. Although there are some guiding principles and some core areas of instructional objectives and practices having wide applicability in all schools, there are also many significant differences in communities relatively close to each other geographically. These differences may justify varied patterns of objectives, of instructional materials and methods, and of behavioral outcomes in the pupils of any two such schools. After the teachers and other school officials have formulated both general and specific objectives and have identified resulting outcomes, the natural next step is the selection or the construction of tests and techniques that will measure the degree to which pupils have attained the desired outcomes. In fact, the degree to which the general outcomes proposed are capable of objective evaluation may well be a significant criterion of their suitability.

Great care must be exercised in the application of the last two of these three methods of curriculum validation of a test, fruitful as they are when appropriately used, if the test is to possess the expected degree of validity. Instructional objectives have often been in the past, and sometimes are even today, stated in such vague and inexact terms that it becomes impossible to obtain a sufficiently precise understanding of their meaning for effective use in test selection or construction. Furthermore, instructional outcomes as envisioned often in the past and even sometimes today may be stated in terms of outcomes expected or desired some years in the future rather than in the near or immediate future. Even the indirect measurement of such remote or ultimate outcomes presents a task for which there is at present no feasible method. Outcomes of an operational and definite type, having meanings readily understandable by all, are most significant for the test selector or test constructor. Similarly, outcomes attainable in the near future, although conceived in terms of later realization of the ultimate objectives, seem most meaningful for the teacher in making or selecting tests. These points are developed more fully in Chapter 5 of this book but they



are mentioned here because of their importance in an approach to test validation.

## Statistical validity

A second method of validating tests is by means of statistical techniques. Methods frequently used involve the determination of the correlation between test scores and such criteria as teachers' marks, ratings of expert judges, scores on other tests designed for the same type of use, and measures of success on certain types of future outcomes. Basic to this method is the belief that the test is valid if high correlations are obtained between scores on it and the criterion measures, and implied is the belief that the criterion measures may be accepted as measurement standards. Correlation coefficients obtained from the types of situations named above are called *coefficients of validity*.

*Correlation with school marks.* The method of validation by correlation with school marks assumes that in the long run a test has validity if the pupils' scores on it are closely related to their achievement in the subject. That is, a test in language must have considerable validity if pupils whose school marks in the subject are consistently high make the superior scores on the tests and if pupils whose school marks in the course are low usually make the inferior scores on the test. In spite of the apparent unreliability of teachers' marks for refined measurements, an educational test that consistently picks out the pupils who, in the teacher's judgment of a specific ability, are superior or inferior, probably does have significant validity.

*Correlation with ratings of expert judges.* This procedure is related in many respects to the one discussed above. To the extent that teachers' marks are the judgments of experts, the two procedures are identical.

*Correlation with other known measures.* This method may be utilized in fields in which extensive critical work in test development has already been done. There would be reason to doubt the validity of a factual achievement test in American history that did not show some relationship to achievement of knowledge outcomes as measured by other valid tests in this subject. This is particularly true in the content subjects as contrasted with the skill subjects. This method of test validation is most frequently used when an outstand-

ingly superior test is available to serve as the criterion. For example, the individual intelligence test constitutes the best basis at present for the validation of group intelligence tests.

*Correlation with measures of future outcomes.* This method of validation is used primarily with prognostic and sometimes with aptitude tests. As the purpose of a prognostic test is to predict future outcomes, e.g., the success of ninth-grade pupils in a course in first-year algebra, the degree to which scores on the test are related to measures of the outcomes the test attempts to predict indicates the validity of the test.

Another group of validation methods primarily statistical in nature but not involving correlation coefficients is based on differences in test scores made by pupils having different subject matter backgrounds or levels of maturity. The two such methods discussed below are used primarily by the maker of standardized tests.

*Accomplishment of widely spaced groups.* One of the readily recognized evidences of validity in test content is the power of such material to reveal significant differences in the accomplishment of widely spaced groups. For example, a performance test for use in the eighth-grade woodworking shop might be validated by administering it to groups of eighth-grade pupils who have had a semester of shopwork, and to similar eighth-grade pupils who have taken no industrial work in this field. If the test is valid in content, the differences in the scores made by the two groups should be significant. It is assumed, of course, that the pupils have actually learned something in the semester course in shopwork. This procedure is frequently used in the validation of aptitude tests and of other tests in which rather highly specialized skills are involved.

*Rise in percentage of success.* This method is based on the changes that education and maturation bring about. A valid reading test is expected to show significant increases in scores indicative of increased achievement as the tests are used in successive school grades. If twelve-year-old children do not demonstrate a higher level of mental maturity than eleven-year-old children on the same test, there is reason to question the validity of the test of mental ability.

*Social utility.* The validation of content in terms of social utility assumes that the course of study itself is based on that point of view. This procedure is distinctly in line with modern theory in curriculum construction. An example of this approach to spelling test construction is the use of words that exhaustive word counts



have shown to be most widely used in written language, and therefore words that the pupils need most to be able to spell correctly. Also, home mechanics tests might be based in part on the skills, such as fixing a leaking water tap, hanging a window weight, or wiring a buzzer, that activity analyses have shown to be most frequently required in the maintenance of household equipment.

## Psychological and logical validity

There are certain subjects in which it appears to be impossible to secure an objective or statistical basis of validation. In general, these subjects are in the complex fields made up of many interrelated abilities as contrasted with those practical skill areas where the tested performance either is an exact representation of, or a very similar substitute for, the instructional outcome sought. Analysis both of the desired outcome and of the proposed test by psychological and logical methods may well reveal a sufficient degree of commonality or of similarity to justify the belief that the test constitutes a valid measure of the outcome. Such methods are followed quite frequently in such complex fields as language and the reading-study skills areas.

## 2 RELIABILITY

A test is said to be reliable when it functions consistently. *The reliability of an examination depends on the efficiency with which a test measures what it does measure.* This statement may appear on the surface to conflict with, or to repeat, the statement in the preceding section concerning the validity of an examination. Such is not the case, however. A test may satisfactorily test what it *does* test without to any effective degree testing what its user *attempts* to test. However, it cannot efficiently measure what it *attempts* to measure unless it efficiently measures whatever it *does* measure. This is equivalent to the statement that a test may be reliable without being valid but that it cannot be valid unless it is reliable. Therefore, *reliability is really an aspect or a phase of validity.*

When a reliable test is used with the type of pupils and for the purpose for which it is intended, it will also be valid. This concept is fundamentally a restatement of the fact brought out in the above section—that validity is *specific* and that it depends not only on test content but also on the proper use of the test. Thus, reliability,

even though it is an aspect of validity, is *general* in nature. Reliability, in turn, has two aspects, adequacy and objectivity. These will be discussed later in this section.

Reliability is most frequently expressed by the use of the coefficient of correlation. In each of the four methods presented below for obtaining or estimating the reliability coefficient, it is the internal consistency or self-consistency of the test that is being evaluated. Only the general methods of obtaining the coefficients and discussions of their applications are given here. The statistical procedures involved in obtaining the various coefficients are presented in Chapter 14.

*Reliability coefficient.* The method of determining the reliability of a test is by means of correlating scores on two equivalent forms of the same test given successively by the same procedure to the same group of pupils. The resulting measure is called the *coefficient of reliability*. Thus, as is true of the validity coefficient, the reliability coefficient is simply a special application of the coefficient of correlation. Students interested in making a critical analysis of the reliabilities of standardized tests should doubtless do so on the basis of the correspondence of scores on two forms of the test. The resultant coefficient is likely to be safe and to be free from the factors making for artificially high relationships that sometimes result from less critical methods.

One method of estimating test reliability when two forms of the test are not available or cannot conveniently be given makes use of the *retesting coefficient*. This coefficient, which is also a special application of the coefficient of correlation, is sometimes used when only one form of a test is available. The test is given to the group of pupils twice under similar testing conditions and the retesting coefficient is the correlation coefficient between the two sets of scores. The second administration of the test should not too quickly follow the first, for a significant increase of scores may result from memory of the previous experience with the test, but neither should it be delayed until forgetting has operated to a high degree. In any event, some increase of scores will probably result from the practice effect. Lindquist pointed out that this method is in general unsatisfactory, especially for achievement tests, and that it results in a spuriously high coefficient.<sup>3</sup>

<sup>3</sup> E. F. Lindquist, *A First Course in Statistics*, Revised edition. Houghton Mifflin Co., Boston, 1942. p. 219-20.



A second method of estimating the reliability of a test is by means of the *chance-half coefficient*. The test is given to a group of pupils and their scores are then obtained for two arbitrarily determined halves of the test. Usual methods of dividing a test into chance-halves are: (1) obtaining separate scores on the odd-numbered and on the even-numbered items, or (2) obtaining separate scores on items 1, 4, 5, 8, 9, 12, 13, etc., and on items 2, 3, 6, 7, 10, 11, etc., to equalize difficulty of the two half-scores when the items are in a scaled order of difficulty. The correlation coefficient obtained between the two sets of scores indicates the degree of conformance between the two chance-halves of the test. The reliability coefficient which would be expected for a test as long as the two halves combined is then found by estimating the correlation by using the *Spearman-Brown Prophecy Formula*.

This method of estimating test reliability has been popular in the past, since it involves a relatively small amount of labor and expense. Lindquist pointed out that the coefficients of reliability estimated by this method are less dependable than those obtained by correlating scores on two forms of a test and are also likely to be spuriously high.<sup>4</sup> Despite that fact, this is one of the most feasible methods for use with informal objective examinations for which ordinarily no second or alternate form is available.

The third method of estimating test reliability furnishes a *footrule coefficient* which may in some cases be an underestimate but which is never an overestimate of the reliability coefficient. Called a "Foot-rule" coefficient because it admittedly is not the most accurate method, it requires the use of only three facts and measures from the test in a simple formula—the arithmetic mean and standard deviation of the scores and the number of items in the test.<sup>5</sup> Because of its simplicity and because it furnishes a result of sufficient accuracy for many purposes, this method is recommended for use by teachers in estimating the reliability of their informal objective examinations. The method of computing this coefficient is presented in Chapter 14.

As has been suggested above, estimates of reliability coefficients often result in spuriously high or low statements of test reliability. The reliability coefficient itself must be based on a known and ap-

<sup>4</sup> *Ibid.* p. 218-19.

<sup>5</sup> G. F. Kuder and M. W. Richardson, "The Theory of the Estimation of Test Reliability," (Formula 21). *Psychometrika*, 2:151-60; September 1937.

propriate range of ages or grade placement of pupils if it is to mean what it purports to mean. Hence, the reliability coefficient is neither an entirely adequate device nor, for that matter, the only method of indicating the internal consistency of a test.

*Standard error of measurement.* The other increasingly popular device by which test reliability can be estimated is the standard error of measurement. This standard error indicates the degree of accuracy existing in the test score obtained for each pupil on a test. Accuracy here does not relate to the type resulting from lack of errors in computing the scores but rather to the magnitude of sampling errors of the type discussed and illustrated in the following section of this chapter. Since the standard error of measurement is not affected by the range of talent of the pupil group on which it is based, as is the reliability coefficient, it is coming to be recognized as a more concrete way of indicating test reliability than is the reliability coefficient. Methods of obtaining this measure of reliability are developed in Chapter 14.

## Adequacy

The careful test maker never assumes that the instrument he has constructed is capable of measuring all of the factual knowledges or skills that a pupil has acquired in a school course. There are too many by-products and incidental learnings to make this possible. Good teaching should never stress a certain restricted body of facts to the exclusion of all other knowledge. When not only factual knowledges and skills but also concepts, understandings, applications, and tastes and preferences are considered, all significant types of instructional outcomes, the task of measuring all of the outcomes from any course, any instructional unit, or even any single class period becomes hopeless. At best, a test is a sample of certain portions of the total behavior which the examiner considers vital to pupil mastery in the field. Just as a grain buyer samples a carload of wheat by taking samples from different places in the car and grading the samples in order to obtain a measure of quality for the whole carload, a test constructor measures the educational attainments of pupils by constructing test items that represent widely the types of pupil outcomes expected and accepts the scores resulting from their use as representative of the pupils' relative achievements for the entire area sampled by the test items. *Adequacy is the degree to which a*



*test samples sufficiently widely into the subject that the resulting scores are representative of relative total performance in the areas measured.*

The diagram of Figure 2 is used to show the effect of sampling on the reliability of test exercises based on a certain limited field

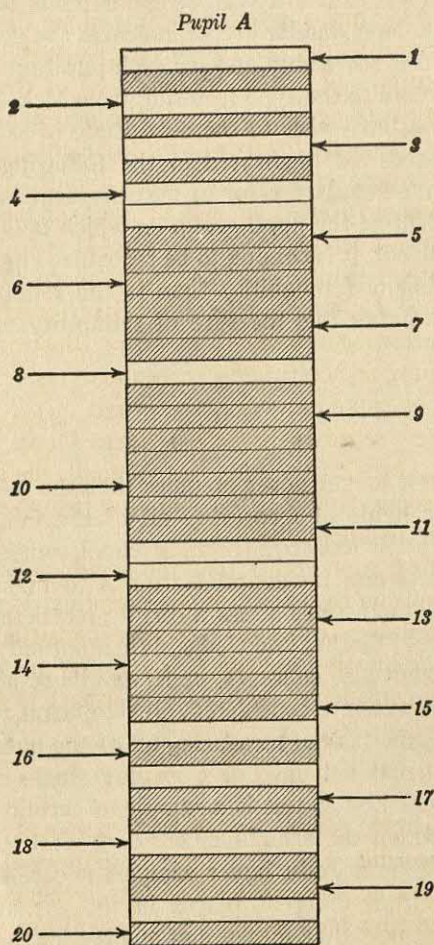


Fig. 2. The principle of sampling

of information. Each of the 40 rectangular spaces in the diagram represents an item of information Pupil A has had an opportunity to learn. The 28 shaded and the 12 unshaded rectangles represent

respectively the items he has and has not mastered. Thus, he has correct information on 28, or 70 per cent, of the 40 items.

If a very limited sampling comprised of items 1 to 5 is taken at one end of the field of information, Pupil A will be equipped to answer correctly only item 5. However, if the last five items, 16 to 20, are selected for the test, he should be able to answer correctly all except item 18. Thus, through sampling alone there is a variation from 20 per cent to 80 per cent of correct responses. Again, if he is separately tested on one test composed of all of the even-numbered items, and another composed of all of the odd-numbered items, he should have right answers for five items and eight items respectively, giving him percentage scores of 50 and 80. Finally, if he is tested on all twenty numbered items, he will answer thirteen correctly and consequently have a percentage score of 65. It is to be noted that as the number of items is increased the pupil's success on the test more nearly approaches the actual amount of his information in this field.

The use of percentage scores in the above illustration is not to be taken as condoning or accepting a percentage basis of marking. The purpose here is to make clear that the extent of the sampling which the items in a test represent is an important factor in the accuracy of the resulting scores. If the sampling is small, the scores are likely to be unfair to some pupils; if the sampling is ample, the scores are likely to be fair to all pupils.

The above illustration shows that the accuracy or consistency of test scores depends on the extent of the sampling. It was pointed out above that a reliable test must measure consistently. Therefore, adequacy of sampling is essential to reliability in a test, and adequacy should be considered as a phase or aspect of reliability.

## Objectivity

A test is objective when the scorer's personal judgment does not affect the scoring. The need for elimination of the subjective factor in the marking of examinations was recognized early in the growth of the testing movement. This recognition was one of the major factors contributing to the development of the standardized and informal objective tests. *Objectivity in a test makes for the elimination of the opinion, bias, or judgment of the person who scores it.*

In general, objective test items are so worded that only one answer



satisfies the requirements of the statement. The distinct advantage of selecting highly objective items for use in educational tests is that there can be little or no disagreement on what is the correct answer. This means that outside of purely chance errors there should be no variation in the scores assigned to a given test by different persons or by the same person on different occasions.

The effect of objective items on the accuracy of marks is shown by an analysis of the marks assigned by a group of ten teachers to two examinations in geography written by the same pupil over the same subject matter. One paper consisted of his answers to 10 essay questions, while the other gave his answers to 40 true-false items. Since each teacher marked the papers independently and at different times, there was little chance for the mark assigned previously to carry over. The range of scores shown in Table 1 for the essay examination was from 76 to 90, the average of the ten scores was 83, and the average amount by which the ten scores deviated from 83 was three score points. On the other hand, one score of 30, one score of 32, and eight scores of 31 were assigned to the true-false test.

TABLE 1. Scores assigned by ten teachers to an essay and a true-false examination over the same material in geography

Teacher	Scores Assigned	
	Essay-type (10 Questions)	True-false (40 Items)
A	84	31
B	80	32
C	88	31
D	76	30
E	82	31
F	90	31
G	85	31
H	82	31
I	83	31
J	80	31
Average . . . . .	83	31
Average error (deviation from the average) . .	3	.2

Thus, the average amount by which the scores deviated from the average score of 31 was but two-tenths of a score point. The relative objectivity of the two types of examinations is shown definitely by these findings.

Objectivity, as well as validity and reliability, of a test may be expressed by the use of the correlation coefficient. The coefficient obtained between scores or marks assigned to a group of papers by the same individual at two different times is sometimes called the *coefficient of objectivity*. However, this coefficient is less widely used than are those for estimating validity and reliability, inasmuch as the fact is quite obvious that the best types of objective test items are relatively high in objectivity.

What was pointed out above as being true for adequacy of sampling is also true for objectivity—both are essential to test reliability and both are therefore aspects or phases, although independent ones, of reliability.

### 3 PRACTICALITY

A good examination must also possess certain characteristics of a quite practical nature. These characteristics of administrability, scorability, and economy have to do with ease of administering the test and scoring the results and with the requirements in labor and financial outlay entailed in using the examination effectively.

#### Administrability

While administrability is not one of the major criteria, it is nevertheless one worthy of much practical consideration. *Administrability is the characteristic of a test that is concerned with ease, clarity, and uniformity in its administration and in preparation for its administration.*

Ease of administration must be evaluated from two distinct points of view—that of the test administrator and that of the pupils taking the test. Specifications should be complete and precise both for advance preparations and for actual test administration. Definite provisions should be made for the preparation, distribution, and collection of test materials, for oral instructions by the examiner preceding, during, and at the end of the examination, for written directions to the pupils covering the test as a whole and for each separate part,



for timing of the test or test parts, and for a variety of other factors.

Instructions to the pupils, whether they are written, oral, or both written and oral, should be simple, clear, and concise. Any unusual types of items or test elements of complex nature should be introduced by sample items and illustrated by practice exercises. The test format should be such that pupils will have no difficulties in reading the items, in recording their answers, in moving from one page to the next or from one part to the next, and in various other practical uses of the testing materials. Illustrations should be clear-cut and easily tied up with the appropriate test items. The page size, length of line, size and style of type, and other mechanical features should be such as to facilitate rather than hamper the administration of the test.

The provision of direct and simple methods of recording, translating, and interpreting the results of the test is another important aspect of ease of administration. Otherwise useful instruments sometimes involve complex processes in turning the raw scores into practical and useful forms.

For a standardized test provisions of these types should be made in the test booklet and manual. For informal objective tests, performance tests, essay tests, and other evaluation procedures and tools great care should also be taken in standardizing the examination procedure by the advance preparation of specifications for administration. Assurance that a test possesses the characteristics of administrability discussed above is best given by adequate printed or written specifications which are made available to, and are followed strictly by, the test administrator.

## Scorability

*The results of a test possessing scorability should be obtainable in as simple, rapid, and routine a manner as is commensurate with their importance.* It is desirable that tests be subject to accurate scoring by clerical workers or other persons not conversant with their content. Various methods of facilitating the scoring of tests, and thereby increasing their scorability, have been devised. Among these methods, discussed in Chapter 5 of this volume, are the use of prepared keys, the use of separate answer sheets to be scored by hand, and the use of separate answer sheets to be scored by machine.

A convenient form of answer key or stencil should be provided for standardized tests, and the manual of directions should carry complete instructions for scoring the instrument. The scoring keys should be arranged so that easy and accurate scoring of the tests can be accomplished. Properly spaced answers on scoring keys for informal objective examinations can be prepared by filling in the correct answers on a copy of the test and converting it into a set of strip keys, cutout stencils, or a combination of the two, according to the nature of the test parts.

## Economy

Economy is certainly not one of the major criteria of a good test, but it is a factor that must be given consideration. Real economy in testing will not be achieved by indiscriminate use of cheap tests or testing methods, but it is equally true that the most costly instruments and methods are not necessarily the best. In the last analysis, *the economy of a testing program should be computed in terms of the validity of the tests per unit of cost.*

There are many devices by which costs of testing can be kept low without reducing the effectiveness of a measurement program. Informal objective tests can be prepared by use of the mimeograph or gelatine plate, and some types may even be given by a blackboard method or orally. The economies of time made possible through the use of some of the scoring devices mentioned in a preceding section of this chapter result in real financial saving. Cooperative testing programs operating under institutional or public educational auspices in many of the states offer testing services to the schools at very low rates. Test booklets that are not necessarily destroyed by one use are now available for many standardized tests, whether machine-scoring or hand-scoring is used. Therefore, an effective testing program need not be dependent on great financial outlay.

## 4 COMPARABILITY

*A test possesses comparability when scores resulting from its use can be interpreted in terms of a common base that has natural or accepted meaning.* There are two means whereby comparability of results is established for standardized tests: (1) availability of



duplicate forms of the test, and (2) availability of adequate norms. Standardized tests should be accompanied in the test manual or elsewhere by adequate tables of norms adapted in type to the age and grade levels for which the test is intended and to the types of abilities it measures. By the use of such norms, individual pupils or class groups can be compared with average performance for pupils of similar age, of similar grade placement, or who are taking the same course. By the use of duplicate forms of a test, results from testing before and after a period of instruction can be made comparable without the necessity of using the same test twice.

Comparability of results can be established for informal objective tests by the simple statistical procedures presented in Chapter 13. In a sense, a series of duplicate forms is established when different class groups are tested over a period of several years, even though the tests used from year to year may overlap considerably in content. In a sense, also, norms can be statistically established on the basis of results from any but very small classes, although such norms do not possess the reliability and wide significance of norms for standardized tests that are based on extensive pupil populations.

The importance of comparability of test results is great, for without comparability of measures some of the major values resulting from the use of tests are lost.

## 5 UTILITY

A test may possess adequately all of the important characteristics of a good test discussed above and yet be of relatively little value for use in a particular school situation. *A test possesses utility to the degree that it satisfactorily serves a definite need in the situation in which it is used.* Unless tests are selected or constructed for definitely conceived purposes and their results used in an intelligent attempt to bring about the desired results, they are of little value and may even, in fact, be harmful. The modern teacher has a definite purpose in mind when he administers tests and makes as effective use as possible of the results in the guidance of his pupils.

If the test is standardized, simple illustrations of the methods of interpreting and using the results should be given in the manual. If the test is one which the teacher constructs, its utility depends largely upon the foresight of the teacher in so planning the test and its use that the results will serve the needs of the local classroom.

Utility may in a sense be considered a final master criterion. It is certainly not entirely distinct from the other criteria, but it may be an effective final check on the value of the test.

### Topics for Discussion

1. What is meant by the validity of an examination? Define or explain in several ways. Is it a general or a specific concept?
2. Discuss and illustrate the two major methods by which validity is obtained in a test. What is the final or ultimate basis on which test validation depends?
3. How does the concept of validity differ for standardized and informal objective tests?
4. What cautions should be observed in the formulation of instructional outcomes as a basis for test validation?
5. Define or explain reliability as a criterion of a good examination. Is it a general or a specific concept?
6. Briefly discuss the methods by which the reliability coefficient of a test can be obtained or estimated. Consider the relative merits of the several methods.
7. Is a valid test necessarily reliable? Explain. Is a reliable test necessarily valid? Explain.
8. Show how test adequacy is essential to reliability. How is adequacy assured?
9. Show how objectivity contributes to reliability. What are the specific features of an objective test?
10. By what means is administrability obtained in a test? Is this an important criterion of a good examination?
11. How may scorability be obtained in an examination?
12. What is the importance of economy as a criterion?
13. What is meant by comparability as a criterion of a good examination? What are the two major means of attaining comparability?
14. Explain how norms or their equivalent are essential to an examination that possesses comparability.
15. In what way is utility in a sense a master criterion of a good examination?
16. Review the criteria of a good examination and show why a good test must be properly balanced in all respects if it is to serve its purpose efficiently.
17. How do the criteria of a good examination apply to other types of measuring instruments and techniques?



## Selected References

- BROWNELL, WILLIAM A. "Some Neglected Criteria for Evaluating Classroom Tests." *Appraising the Elementary School Program*. Sixteenth Yearbook of the Department of Elementary School Principals. Washington, D. C.: National Education Association, 1937. p. 485-92.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. New Brunswick, N. J.: Mental Measurements Yearbook, 1941.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949.
- CURETON, EDWARD E. "Validity." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 16.
- ENGELHART, MAX D. "Examinations." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 407-14.
- FLANAGAN, JOHN C. "Units, Scales, and Norms." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 17.
- GERBERICH, J. RAYMOND, AND PETERS, CHARLES C. "Reliability." *Encyclopedia of Modern Education*. New York: Philosophical Library, 1943. p. 673-75.
- GREENE, EDWARD B. *Measurements of Human Behavior*. Revised edition. New York: Odyssey Press, 1952. Chapter 3.
- GUILFORD, J. P. "New Standards for Test Evaluation." *Educational and Psychological Measurement*, 6:427-38; Winter 1946.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. Chapter 2.
- LINDQUIST, E. F. "The Theory of Test Construction." *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Co., 1936. Chapter 2.
- MICHEELS, WILLIAM J., AND KARNES, M. RAY. *Measuring Educational Achievement*. New York: McGraw-Hill Book Co., Inc., 1950. Chapter 4.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. Chapter 12.
- ROSS, C. C. *Measurement in Today's Schools*. Second edition. New York: Prentice-Hall, Inc., 1947. Chapter 3.
- SCATES, DOUGLAS E. "Differences between Measurement Criteria of

Pure Scientists and of Classroom Teachers." *Journal of Educational Research*, 37:1-13; September 1943.

SPAULDING, GERALDINE. "Reproducing the Test." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 11.

THORNDIKE, ROBERT L. "Reliability." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 15.

THUT, I. N., AND GERBERICH, J. RAYMOND. *Foundations of Method for Secondary Schools*. New York: McGraw-Hill Book Co., Inc., 1949. p. 101-7.

TRAVERS, ROBERT M. W. *How To Make Achievement Tests*. New York: Odyssey Press, 1950. Chapter 7.

TRAXLER, ARTHUR E. "Administering and Scoring the Objective Test." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 10.



## ***Constructing and Using Standardized Tests***

THE FOLLOWING PROBLEMS in the construction and use of standardized tests are considered in this chapter:

- A. Meaning of the standardization process.
- B. Controlling validity and difficulty of test items.
- C. Equating test forms.
- D. Deriving norms for standardized tests.
- E. Establishing final validity and reliability of tests.
- F. Using standardized tests in instruction, guidance, supervision, and administration.
- G. Instructional uses of educational tests.
- H. Diagnosis and analysis as related to remedial instruction.
- I. Planning the testing program.
- J. Selecting and administering the tests.
- K. Securing and using the results in the classroom.

### **1 CONSTRUCTING STANDARDIZED TESTS**

A treatment of the problems of constructing and refining standardized tests that would be sufficiently detailed to afford an adequate guide to the inexperienced worker who might have ambitions to construct such a test would run far beyond the confines of a single chapter in this textbook. It would be a volume in itself. This chapter, however, should be adequate to make the student or the classroom teacher more critical of all types of standardized measuring instru-

ments, and at the same time more appreciative of the technical skill and expense in time and money required to produce a commercial educational test of a quality adequate to stand up under present-day criteria.

## Meaning of standardization

Standardization, or the process of deriving comparative norms, is quite frequently designated as the single factor that distinguishes the formal standardized test from the informal objective test. However, a program of standardization demands a more critical analysis of subject matter, a more careful formulation of exercise material, a more exacting refinement of the techniques of evaluating test items, more critical standards of equality of items and of test forms, and more rigid statistical analysis than are usual for the informal objective test. Thus real differences in the two types of tests appear. It becomes clear that the mere derivation of a set of norms for a test does not in itself make the instrument a standardized test. The matter of securing norms that facilitate interpretation of the test results is undoubtedly the most important phase of test standardization, but it is only one of several important procedures that are closely related to the standardization process.

## Establishing validity of test content

The maker of the standardized test faces the rather complicated problem of preparing an examination over content and for a group of pupils he has not specifically taught. To be reasonably certain that he is fair in the selection of items, he must take care to include only those aspects of the course that are very likely to receive instructional emphasis. Naturally this requires at the outset the selection of test content that must be general enough to fit into any school situation in which the course is taught.

The difficulties encountered in the selection of item content for the standardized test depend to a certain extent upon the nature of the skills, knowledges, concepts, understandings, or applications to be tested. If the field is one in which the objectives and outcomes are clean-cut and readily identified, the problem may be comparatively simple. In the case of arithmetic, a subject in which the fundamental facts and skills are well known, the selection of content suitable for



use in a standardized test is a relatively simple matter. Many tests of acceptable validity are available in this field. Exactly the reverse is true in certain content courses. The fields in which the instructional aims are specific or highly factual lend themselves readily to the construction of standardized tests. Tests in fields in which the knowledges, skills, attitudes, and other outcomes are of a more indefinite nature are much more difficult to validate.

In most subjects the validity of the test content is very difficult to establish by acceptable statistical or other objective means. In certain fields it is practically impossible. In one subject a certain validation procedure may be effective and acceptable; in another it may be completely unsuited for use. Makers of standardized tests have resorted to many different types of validation procedures, some of which are discussed in a later chapter.

### Constructing and validating test items

The discussion of the problems of constructing and using informal objective tests in Chapter 7 points out certain principles that are to be observed in the selection of the item content for any type of objective test, whether standardized or informal objective. Therefore, methods used by the makers of standardized tests in the selection of objective item types to use for each fact, principle, relationship, or outcome which they wish to test and in the actual construction of various item types are discussed only briefly here.

Test validity depends upon (1) the validity of the content in general and (2) the validity of the individual items of which the test consists. The validity of the items before they have been tried out with a group of pupils depends upon the ability of the test constructor to select the objective item form best suited for each outcome to be tested as well as so to construct the item that it measures the desired type of pupil behavior and has none of the weaknesses pointed out in the following pages of this chapter. Objective evidence concerning item validities, however, is secured only by (1) the actual administration of the test in preliminary form to a large group of typical pupils and (2) a detailed statistical analysis of the results item by item. Inasmuch as many of the original items may not be satisfactory, and hence must be discarded or revised, the preliminary form of the test frequently contains many more items than will appear in the final forms.

*Objectivity.* Objectivity in the test item is such an important element in the reliability of measurement which the test affords that it would be difficult to conceive of a standardized test made up of items not characterized by the quality known as objectivity. In general, objectivity is determined by the form in which the test item is stated. The builder of a standardized educational test faces the very difficult problem of determining the precise form of objective technique that best fits the subject he wishes to test. In most cases this is a problem that can be answered only by experimentation and as the result of experience.

After the instructional areas to be covered by the standardized test have been determined, the test maker must proceed to analyze the subject matter into elements representing the basic concepts. These important elements may then be stated in some objective form, the form selected depending to a certain degree upon the nature of the objectives and outcomes of the course and somewhat on the maturity level of the pupils with whom it is to be used. Frequently it is necessary to prepare certain of these basic concepts in two or three different objective forms, selecting for final use the types that perform best under experimental conditions. In addition to the item type to which the content to be tested is best adapted, there are three other factors that characterize the objective test item. These are briefly discussed below.

1. *Uniformity of response.* Test items which otherwise appear to meet the requirements of objectivity frequently are so stated that they allow considerable variation in response. This weakness in the test items is more likely to be found in recall or completion items than in any other forms. Since items of these types are rarely used in standardized tests today, it may be sufficient to point out here that, other things being equal, items that set up conditions encouraging a multiplicity of answers should be eliminated or made more objective.

2. *Sparing use of clues and suggestions.* One of the common criticisms of the objective test item is that it contains many suggestive elements or clues that the pupils soon learn to recognize as indicating a certain type of response. Unquestionably this is a significant criticism of many objective forms. In formulating objective items, great care should be taken to see that undue suggestion of the truth or falsity of an item is not inherent in the form of the statement, although this is frequently very difficult to avoid. Some



experience in the making of tests of the alternate-response types indicates that the false or negative statements are more difficult to formulate and are more likely to contain recognizable clues. The systematic use of such terms as *always*, *never*, *no*, *not*, as well as such prefixes as *un-* and *in-* in the negative forms of items may easily lead the pupil to spot them at once as clues.

3. *Freedom from ambiguity.* The elimination of ambiguity, or the possibility of misinterpretation, is one of the most difficult problems the test maker has to face. To keep ambiguity out of an item frequently means that he must simplify the statement. When the concept itself is simple, it is almost impossible to escape making the statement so obvious that its validity as a test item is reduced. Another aspect of the problem of ambiguity in test items is the fact that there are certain items which to the ignorant or poorly informed pupil are perfectly straightforward and clear but which to the critical and well-informed pupil involve implications that cloud the issue. That is, the better the pupil is informed in the field represented by the item the more likely he is to be confused by it. This is a phase of ambiguity in the statement of items that is closely tied up with item difficulty. This point is discussed below as one of the major criteria for the selection of test items.

*Difficulty.* The difficulty of a test item is usually expressed in terms of the number or percentage of pupils of a certain classification who respond to it correctly. The determination of the optimum difficulty of the test items to be used in a standardized test is a problem on which there is not complete agreement among test specialists. Some test authorities prefer approximately equal numbers of items at all levels from very easy to very difficult, while others prefer to use a few easy and a few difficult items but to have the majority near the 50 per cent difficulty level. They are in general agreement, however, that the test as a whole should have about 50 per cent difficulty for the average pupil.

The common practice in test construction is to attempt to prepare items covering a wide range of difficulty, from very easy to very difficult. *Items are not suitable for inclusion in the test if they are so easy that no pupil of the type on which the tests are to be used fails to respond correctly.* The presence of such items would merely serve to lengthen the test without adding to the reliability of its measurement. In a similar way, *items that are so difficult that no pupil is able to respond correctly should not be included in the test.*

Thus, items that lie at the extremes of difficulty, 100 per cent failure and 100 per cent success, are useless, since no one is able to tell how far beyond these limits the difficulties may lie. An item that does not in a very direct way serve to differentiate between pupils at various levels of achievement has no place in an educational test, since it adds only useless dead weight.

Modern practice in the arrangement of standardized test items tends to follow the procedure of presenting items covering a wide range of difficulty in ascending order from the very easy to the most difficult. This plan makes it possible for the lower-grade or less able pupils to respond to certain items within their level of mastery without being unduly discouraged by being confronted at the outset with exercises of prohibitive difficulty. On the other hand, it also causes the more able pupils to waste a certain amount of time working through a large number of items that are time-consuming but are not hard enough to bring out their real abilities. The allowance of liberal working periods for such tests tends to take care of this difficulty somewhat. Thus each pupil is allowed to work long enough to reach the level at which his abilities are taxed to the utmost. If the test items are carefully scaled in such a way that there is a gradual and continuous rise in difficulty, a relatively small amount of time is lost by the superior pupils in working on items that are far below their abilities. Similarly, the levels of ability of the less accomplished are revealed quite promptly and accurately. The problems involved in the scaling or statistical evaluation of test items are rather technical and require a much more extensive treatment than can be justified in this volume.

*Discriminative power.* The basic function of all educational measurement is to place individuals along a defined scale in accordance with differences in their achievement. Such a function implies high discriminative power on the part of the test. Since tests are made up of separate items, it is clear that each item comprising a test must have this quality in a maximum degree if the total test is to possess it.

Discriminative power in a test or a test item means that a different quality or magnitude of response may be expected from individuals or groups possessing the abilities in question in varying degree. Pupils with superior ability should answer the item correctly more often than should inferior pupils. This suggests a method by which the power of a test item to discriminate or distinguish between



groups of pupils may be determined. The practical implications of this procedure may be illustrated quite simply.

An experimental test has been given to a class of 100 pupils having the normal range of ability in the subject. The tests have been corrected by the use of the answer key, and the score of each pupil in terms of the number of items answered correctly has been determined. On the basis of these scores, the class of 100 pupils may be divided into three groups. The 27 per cent of the pupils making the highest scores constitute the superior group; the 27 per cent making the lowest scores comprise the inferior group. The 46 per cent of the class in the middle are not considered in computing this index of discrimination. The use of the 27 per cent comprising each of the extremes follows a proposal made by Kelley and further exploited by Flanagan<sup>1</sup> for this purpose. The next step involves an item count for all of the items in the test, showing the number and percentage of pupils in the superior group compared with similar data for the inferior group. A summary of a brief sampling of items from a typical test is given in Table 2.

TABLE 2. Discriminative power of test items in percentages of success by superior and inferior groups

Item	Superior Group High 27%	Inferior Group Low 27%	Index of Discrimination
I	12	4	.23
12	6	4	.08
23	8	14	— .13
44	10	18	— .15
55	24	13	.15
76	42	22	.23
97	52	12	.46
108	80	36	.46
129	90	86	.08
140	92	40	.59

This table indicates that Item 1 was answered correctly by 12 per cent of the superior and by 4 per cent of the inferior pupils. This

<sup>1</sup> John C. Flanagan, "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from the Data at the Tails of the Distributions." *Journal of Educational Psychology*, 30:674-80; December 1939.

item thus shows great difficulty and a limited power to discriminate between good and poor achievement. The fact that the item is answered correctly by such a small proportion of all pupils (average 8 per cent) indicates that its difficulty is great. Item 44, however, is correctly answered by a smaller percentage of superior pupils than of inferior pupils. This is shown by the negative discrimination index of  $-.15$ . This shows that the item is at fault or the wrong facts have been taught in this subject. The item should probably be eliminated from the test. Items 97, 108, and 140, with positive indexes of .46, .46, and .59, are probably good enough to retain in the test.

This method of determining the discriminative power of test items is widely used in the critical analysis of test items for standardized tests. The classroom teacher who is interested in the experimental development and analysis of informal objective examinations will also find in the method illustrated a very satisfactory procedure for determining the quality of test items.

### Methods of equating test forms

Two or more forms of an educational test are considered to be equated when practically identical scores on each are made by the same individuals or by individuals of the same ability. This means that the forms of the test must be made up of test items that parallel one another closely in difficulty. In practice, such close equality of item difficulty in alternate forms is obtained in one of three ways.

1. The first procedure involves the preparation of large numbers of items covering the total range of the outcomes to be tested, on the chance that there will be a sufficient number of items at each of many difficulty levels to permit the pairing of items of equivalent difficulty in the alternate forms of the test. When this is done, the alternate forms of the test may be considered roughly equal in difficulty, but there will be only a very general and limited equivalence of content.

2. The second procedure involves the preparation of parallel items on certain selected, important concepts. One item may test the identification of the concept, while the other may test the identification of an additional phase of the concept or some phase of the identification of the procedure involved.



3. A third procedure that permits the establishment of comparable forms of tests by the use of derived scores is mentioned here, although the complexity of the statistical techniques necessary and the variety of derived scores which are used in this way make a complete presentation impracticable at this point. It may suffice here to say that the derived scores are so established that they have constant meanings, whether or not they are obtained on the same form of the test or from the same pupil group, and that the method of establishing a "normalized group" is basic to the procedure. Several of the most widely used derived scores that are in general based on this type of procedure are presented in Chapter 13.

### Deriving test norms

Norms provide the user of a standardized test with the basis for a practical interpretation and application of the results. Unless the norms that accompany a test reflect an accurate picture of typical accomplishment, they are useless and they render the test itself useless.

Early in the history of objective testing, practically all that the development of a standardized test required was to give a few test exercises to a hundred or more pupils in different school systems. The results were then compiled and submitted as norms. The standardized test differed from a reasonably good informal objective test mainly in the fact that the former had been tried out with more pupils in a larger number of different classes. In fact, many informal examinations of the objective type meet all criteria of standard tests except that of having norms for the evaluation of their scores. However, test standardization as it is now interpreted means much more than the mere derivation of norms, although the existence of norms is still the most distinctive feature of the standardized test.

Norms are tables of information necessary for the interpretation of test scores and are obtained by giving the particular test to a large and representative sampling of pupils in the same grades and of a type similar to the groups with which teachers will use the tests. To the extent that the sampling used in obtaining the norms was distributed over a large population in typical school situations and that the conditions under which the tests are to be administered are rigidly followed by the teachers using the tests, the norms furnish a reliable and useful basis for interpretation.

*Types of norms.* The form in which the norms for a test are provided depends to a large degree upon the level in the school system at which the test is used. The norms are also conditioned somewhat by the nature of the test itself. Tests that are designed for use in the elementary-school grades are usually accompanied by age norms and grade norms and sometimes percentile norms based on grade placement. Tests intended for use in the secondary school are more frequently provided with percentile and grade norms only. Age norms do not seem to be particularly useful at the high-school and college levels, since so many factors other than age operate to affect achievement. Then, too, the curve of growth appears to flatten out quite rapidly after the sixteenth or seventeenth year, so that the increments of growth in achievement from age to age at the upper levels are relatively small.

Brief discussions of grade norms, age norms, percentile grade or subject norms, and percentile norms for schools are given below. The brief illustration of how such norms are derived indicates roughly the procedure used by test makers in establishing norms for a test.

1. *Grade norms.* The grade norms established for most of the commonly used achievement tests are based on the median scores obtained by giving the tests to large groups of pupils within each grade. Such norms provide a reasonably practical basis for the interpretation of class scores as well as of individual accomplishment. In the derivation of grade norms for standardized tests it is a common but not universal practice to express the norms in terms of end-of-the-year achievement. In any event, it is desirable to have the norms clearly indicate the period they are designed to cover. Grade norms thus provide a convenient means of expressing the approximate progress of the pupil through the grades by turning his raw score or standard score into a grade-equivalent score. For example, if the seventh-grade end-of-the-year norm for a certain test were 120 points, and the eighth-grade end-of-the-year norm 140 points, a score of 130 points would be treated as representing achievement halfway through the seventh grade, or a 7.5 grade equivalent.

In many of the modern analytical tests composed of several parts, raw scores frequently are changed into standard scores before the grade norms are established. The data in Table 3, which is an abbreviated table of grade norms taken from the manual for the *Iowa Language Abilities Test*, show the grade equivalents corresponding to standard scores for each subtest and the median standard



scores for this test. Raw scores on each part of this test are changed directly into standard scores as the scoring of each part is completed. The total score on all parts of the test is represented by the median of the several standard scores. In this table a standard score of 160 on Test 1, Spelling, represents a grade equivalent of 8.7, while a similar score on Test 3, Language Usage, is assigned a grade equivalent of 8.3. A median standard score of 160 for the entire test gives the pupil an 8.4 grade equivalent.

2. *Age norms.* At the elementary-school level, age norms of one type or another appear to provide a more adequate basis for the interpretation of individual pupil accomplishment than is possible with grade norms or percentile grade norms alone. The problem of establishing age norms has been complicated by a number of factors arising out of the generally inadequate child accounting practices in the schools and the indifferent attention to the significance of *ageness* as a factor in school progress and accomplishment. In its simplest form the preparation of age norms involves the regrouping of all pupils used in the grade tabulation into chronological age groups regardless of grade location or school progress. The test scores of these chronological age groups are then tabulated, and the means or medians computed. These results are then used as the basis for setting up tables of the scores corresponding to the several age groups. It is readily apparent, however, that many factors other than age are operating to influence the average achievement of pupils grouped in grades. Such factors as over-*ageness*, retardation, and a serious lack of balance between retardation and acceleration are all present. It is found, for example, that while the average chronological age of a seventh-grade pupil might be 13 years and 6 months at the end of the school year, the average test score of pupils of 13 years and 6 months is not at all the same as the end-of-the-year score for the seventh grade. The actual achievement of the under-age pupils is significantly superior to that of over-age groups in a given grade. For the makers of standardized tests the useful implication from this fact is that it makes very apparent the need for norms that will take into account wide differences in maturity, mental ability, or school progress within the grade. While many reputable standardized tests are accompanied by age norms determined without regard to the grade level at which the accomplishment takes place, it appears obvious that interpretations of individual pupil accomplishment on the basis of these norms are likely to be misleading.

TABLE 3. Grade equivalents corresponding to each subtest standard score and the median standard score for the Iowa Language Abilities Test <sup>2</sup>

Standard Score	Spelling Test 1	Word Meaning Test 2	Language Usage Test 3	Gram. Form Recognition Test 4—Int.	Capitalization Test 4—Elem. Test 6—Int.	Punctuation Test 5—Elem. Test 7—Int.	Sentence Sense Test 5—Int.	Median Stand. Score	Standard Score
182	13.1								182
181	12.8								181
180	12.5	13.0							180
179	12.3	12.6		13.0					179
178	12.0	12.3		12.6					178
177	11.8	12.0		12.2					177
176	11.6	11.7		11.8			13.0	13.0 +	176
175	11.3	11.4		11.5	12.2		12.6	12.6	175
174	11.1	11.1	12.8	11.2	11.4		12.1	12.0	174
173	11.0	10.9	11.6	10.9	10.9		11.7	11.5	173
172	10.8	10.7	11.1	10.6	10.5	12.7	11.4	11.1	172
171	10.6	10.5	10.7	10.4	10.2	11.9	11.1	10.7	171
170	10.4	10.3	10.4	10.2	9.9	11.4	10.8	10.4	170
169	10.2	10.1	10.1	9.9	9.7	10.9	10.5	10.1	169
168	10.0	9.9	9.8	9.7	9.5	10.5	10.2	9.9	168
167	9.9	9.7	9.6	9.5	9.3	10.1	10.0	9.7	167
166	9.7	9.5	9.4	9.3	9.1	9.8	9.8	9.5	166
165	9.5	9.3	9.2	9.2	8.9	9.6	9.6	9.3	165
164	9.4	9.1	9.0	9.0	8.7	9.4	9.4	9.1	164
163	9.2	9.0	8.8	8.9	8.6	9.2	9.2	8.9	163
162	9.0	8.8	8.6	8.7	8.4	9.0	9.0	8.7	162
161	8.9	8.6	8.4	8.6	8.3	8.8	8.8	8.6	161
160	8.7	8.5	8.3	8.4	8.1	8.6	8.6	8.4	160
159	8.6	8.3	8.1	8.3	8.0	8.5	8.4	8.3	159
158	8.4	8.2	8.0	8.1	7.8	8.3	8.3	8.1	158
157	8.3	8.0	7.8	8.0	7.7	8.2	8.1	8.0	157
156	8.1	7.9	7.7	7.9	7.6	8.0	7.9	7.8	156
155	8.0	7.8	7.5	7.7	7.5	7.9	7.8	7.7	155
154	7.8	7.6	7.4	7.6	7.3	7.7	7.6	7.6	154
153	7.7	7.5	7.2	7.5	7.2	7.6	7.5	7.4	153
152	7.5	7.3	7.1	7.3	7.1	7.5	7.3	7.3	152
151	7.4	7.2	7.0	7.2	7.0	7.4	7.1	7.2	151
150	7.3	7.1	6.8	7.1	6.9	7.2	7.0	7.1	150
149	7.1	7.0	6.7	6.9	6.8	7.1	6.8	6.9	149
148	7.0	6.8	6.6	6.8	6.7	7.0	6.7	6.8	148
147	6.9	6.7	6.5	6.7	6.6	6.9	6.5	6.7	147
146	6.7	6.6	6.4	6.6	6.4	6.8	6.4	6.6	146
145	6.6	6.4	6.2	6.4	6.3	6.7	6.3	6.5	145
144	6.5	6.3	6.1	6.3	6.2	6.6	6.1	6.4	144
143	6.4	6.2	6.0	6.2	6.1	6.4	6.0	6.2	143
142	6.2	6.1	5.9	6.0	6.0	6.3	5.8	6.1	142
141	6.1	6.0	5.8	5.9	5.9	6.2	5.7	6.0	141
140	6.0	5.9	5.6	5.8	5.8	6.1	5.6	5.9	140
139	5.9	5.8	5.5	5.7	5.7	6.0	5.4	5.8	139
138	5.8	5.7	5.4	5.6	5.6	5.9	5.3	5.7	138
137	5.6	5.5	5.3	5.4	5.5	5.8	5.2	5.6	137
136	5.5	5.4	5.2	5.3	5.4	5.7	5.0	5.5	136
135	5.4	5.3	5.1	5.2	5.3	5.6	4.9	5.4	135

<sup>2</sup> H. A. Greene and H. L. Ballenger, *Directions for Administering: Iowa Language Abilities Test*, Intermediate. World Book Co., Yonkers, N. Y., 1948. Table 10.



3. *Age-at-grade norms.* In the ordinary process of test standardization the establishment of age-at-grade norms involves a number of difficulties having to do with (1) availability of pupil population and (2) statistical procedures resulting from inadequate population groups. While the number of sixth-grade pupils who are between ten and eleven years of age would represent a large portion of the normal sixth-grade population, the number of under-age individuals in the sixth grade who would be nine, eight, or seven years old, and the number of over-age pupils of eleven, twelve, thirteen, or fourteen years of age would fall off very rapidly. Thus, in order to secure reliable age-at-grade norms for all ages within the grade very large numbers of individuals must be tested in order to secure adequate populations in the fringe areas. Otherwise, estimations by extrapolation must be made. While not all of the assumptions on which such estimations are based can be definitely established in practice, it is probable that the injustice done by such estimations is much less serious than would be the failure to provide such differential norms in the first place.

An example of norms given in both age and grade equivalents within the grade is shown in Table 4 for Test C, Language, of the *Iowa Basic Skills Tests*, Advanced. In the interpretation of this test, raw point scores are turned directly into grade equivalents. Thus in this table the first digit of the two-place numbers in the grade columns are to read as the grade location, and the second digit as the proportion of the pupil's progress toward the next grade. For example, the typical second-semester fifth-grade pupil with a chronological age of 13 years and 0 months has a grade equivalent expectancy of 4.3, or 4.3, according to this table. It is possible to determine quickly from this table the grade-equivalent score expectancy for the various age groups.

It is interesting to note here that as almost invariably happens the younger pupils within the grade make the higher grade-equivalent scores up to a certain point, and the older pupils make the lower scores. For example, the table shows that a first-semester sixth-grade pupil who is 13 years and 6 months old has a grade-equivalent score of 4.8. On the other hand a child in the same grade but chronologically three years younger (10 years, 6 months) has a grade equivalent of 6.3. While the typical thirteen-and-a-half-year-old in the second semester of the sixth grade has a grade equivalent of 5.3, a typical child of that age would have a grade equivalent of 10.1 if he were

in the second semester of the ninth grade. It must be apparent that age-at-grade norms offer very useful means for the interpretation of individual pupil accomplishment, especially in the case of pupils who for one reason or another may be over-age or under-age for the grade.

TABLE 4. Age-at-grade norms for the total language score on the Iowa Basic Skills Tests <sup>3</sup>

Grade Equivalents for Grade and Semester (Beginning)										
Age										
Yr. Mo.		5-2	6-1	6-2	7-1	7-2	8-1	8-2	9-1	9-2
17	0	..	..	..	..	..	..	..	..	69
16	9	..	..	..	..	..	..	..	..	70
16	6	..	..	..	..	..	..	..	71	72
16	3	..	..	..	..	..	..	..	72	73
16	0	..	..	..	..	..	..	66	72	74
15	9	..	..	..	..	..	..	67	74	77
15	6	..	..	..	..	..	62	68	76	80
15	3	..	..	..	..	..	63	69	78	86
15	0	..	..	..	..	57	64	71	79	89
14	9	..	..	..	..	58	65	74	83	94
14	6	..	..	..	54	58	66	74	89	98
14	3	..	..	..	56	60	68	76	92	101
14	0	..	..	51	56	62	70	83	93	102
13	9	..	..	52	57	64	73	87	94	102
13	6	..	48	53	59	66	78	88	94	101
13	3	..	48	53	61	68	83	89	96	...
13	0	43	48	54	62	73	83	90	94	...
12	9	44	49	56	65	77	84	90	..	...
12	6	44	52	58	70	78	85	89	..	...
12	3	45	52	60	73	79	86	..	..	...
12	0	47	53	65	73	79	84	..	..	...
11	9	47	57	67	74	80	..	..	..	...
11	6	48	61	68	74	77	..	..	..	...
11	3	52	64	69	74	..	..	..	..	...
11	0	56	64	68	73	..	..	..	..	...
10	9	57	64	68	..	..	..	..	..	...
10	6	58	63	67	..	..	..	..	..	...
10	3	57	63	..	..	..	..	..	..	...
10	0	57	62	..	..	..	..	..	..	...
9	9	57	..	..	..	..	..	..	..	...
9	6	56	..	..	..	..	..	..	..	...

4. *Percentile grade or subject norms.* Relative accomplishment of individual pupils within a grade or who are taking a certain course may also be shown very clearly by turning raw or standard scores

<sup>3</sup> *Manual of General Information: Iowa Every-Pupil Tests of Basic Skills*, Advanced. Houghton Mifflin Co., Boston, 1947.



into percentile scores. This is done by computing the percentile values from the frequency tables for each grade or course distribution and assigning percentile equivalents for each score. Percentile norm tables show for a wide sampling of pupils in a certain grade or course (1) the percentage of pupils exceeding each score or each of a number of equally spaced scores, or (2) the score below which certain percentages of pupils fall. Although percentile norms are customarily presented in one or the other of these methods, there is a great variety in the form of such tables. Percentile scores corresponding to specific raw or standard scores may be reported by grades, by test parts and totals, or only the raw score or standard score equivalents for specified percentiles, quartiles, and deciles may be shown in more compact tables. Whenever percentile grade norms are provided it is recommended that these values be used in the interpretation of individual pupil scores rather than the grade equivalents from the usual grade norms. The overlapping of distributions of scores is so great from grade to grade on most standardized elementary-school tests, and the differences between successive grade medians are often so slight, that grade equivalents may exaggerate differences and lead to unsound interpretations. A sixth-grade pupil assigned a grade equivalent of 8.5 does not belong in the eighth grade. It may be much more accurate to use percentile scores to describe his accomplishment as superior in relation to other sixth-grade pupils.

5. *Percentile norms for school averages.* In large city school systems and in schools participating in state testing programs it frequently becomes desirable to interpret the results in terms of school averages. A comparison of one school with another in the same city or one system with another is not possible through the use of the norms based on individual pupil scores due to the fact that the variability of individual scores is so much greater than the variability of school averages. Percentile norms for school averages are obtained in much the same manner as other percentile norms are derived, except that averages are substituted for individual scores in the grade distributions.

Table 5 illustrates and gives the percentile norms by grades for the school and building averages obtained for total scores on Test C Language, of the *Iowa Basic Skills Test*, Advanced.

The values 53.5, 66.0, and 75.6 given for grades 3, 4, and 5 as the 99th percentiles are all to be read as grade equivalents with the

TABLE 5. Percentile norms for school averages on the total language score of the Iowa Basic Skills Tests <sup>4</sup>

%ile	Grade Equivalents						
	3	4	5	6	7	8	9
99	53.5	66.0	75.6	90.0	99.8	103.0	107.0
95	49.5	60.8	69.6	84.2	95.7	100.8	105.5
90	47.0	58.0	66.0	80.0	91.3	98.4	104.5
85	45.5	56.5	64.5	77.2	88.3	96.4	103.2
80	44.3	55.2	63.4	75.6	86.5	94.8	101.6
75	43.4	54.2	62.6	74.4	85.1	93.3	99.7
70	42.6	53.2	61.8	73.4	83.7	91.9	97.4
65	41.9	52.3	61.0	72.4	82.5	90.6	95.6
60	41.3	51.4	60.2	71.4	81.4	89.3	94.2
55	40.6	50.5	59.3	70.5	80.2	88.0	92.8
50	40.0	49.6	58.4	69.5	79.1	86.8	91.6
45	39.3	48.8	57.5	68.6	77.9	85.5	90.2
40	38.6	47.8	56.6	67.6	76.7	84.3	88.7
35	37.9	46.8	55.6	66.5	75.6	83.1	87.1
30	37.2	45.9	54.5	65.3	74.4	81.9	85.3
25	36.4	44.9	53.5	64.1	73.2	80.6	83.3
20	35.5	43.9	52.4	62.8	71.8	79.2	81.4
15	34.4	42.8	51.0	61.2	70.2	77.7	79.4
10	33.1	41.6	49.4	59.2	68.2	75.9	77.1
5	31.6	40.0	47.2	56.3	65.0	73.3	74.4
1	28.5	36.0	43.3	51.8	69.0	67.0	69.0

decimal points moved one place to the left. The second decimal place is reported here due to its special significance in these school averages.

It should be noted that the type of percentile norms illustrated in Table 5 is designed for use in the interpretation of school results and should not be used in interpreting individual pupil scores. To confuse

<sup>4</sup> Adapted from a series of tables in the *Manual of General Information: Iowa Every-Pupil Tests of Basic Skills*, Advanced. Houghton Mifflin Co., Boston, 1947.



these norms for school averages with percentile grade norms used in interpreting individual pupil achievement would introduce a serious error in test interpretation.

## Norms vs. standards

The use of the term *standardized* in the discussion of tests of the type for which norms are provided has led to the development of a careless tendency to use the words "standards" and "norms" as synonyms. The process of securing the data for the critical analysis of tests and the derivation of suitable norms is properly known as *standardizing*. However, *the term "standard," when used to refer to a level of pupil achievement, implies an ultimate goal to be achieved.* These standards may not actually be reached by any individual, but they are levels of achievement toward which to strive. *Norms are the levels of achievement which typical pupils actually attain.* When considered in the light of these definitions, it is clear that there are few tests which are accompanied by standards. It might be more nearly the truth to call the process of securing these comparative scores known as "norms" by the more descriptive name of "normalizing."

Possibly one of the best illustrations of the differences between standards and norms is to be found in the field of arithmetical computations. The standard of arithmetical accuracy is naturally 100 per cent, for most such computations containing error are useless. However, the actual norm of arithmetical accuracy of computation on a well-known test is from 65 to 70 per cent for the junior high-school grades. That is to say, pupils of these grades work these particular kinds of arithmetical examples with an accuracy of from 65 to 70 per cent instead of the desired ultimate goal of 100 per cent accuracy.

It should be recognized that a norm does not necessarily represent a satisfactory level of achievement. This is particularly true of schools in which instruction and classroom environments are superior and in which pupils, largely because of their satisfactory home environments and heredity, have superior abilities. In any event, teachers should encourage pupils to make the most of their abilities and to surpass test norms whenever they can. Even when a class has average performance that is just at the norm on a test, representing only the attainment of what is expected from a typical class,

approximately half of the pupils will still be below the norm of achievement.

Standards are of two general types. In the first place, there are certain standards of achievement, or minimum essentials, which have been fairly generally accepted by school people for such abilities as handwriting and, in less objective form, reading, spelling, and arithmetic. Although these standards are usually based on the results of standardized testing, and may make use of norm tables for their establishment, they frequently are conceived of as representing the minimum quality and perhaps speed of performance that will adequately equip the pupil for post-school life. For example, the widely accepted standard in handwriting is a quality of 60 on the Ayres' *Scale for Measuring the Handwriting of School Children* at the rate of 70 letters per minute. Although the quality of 60 at the given rate on this scale is approximately the norm for pupils completing the sixth grade, it is also thought of as the standard or minimum ability that should be attained by all pupils before they finish school.

In the second place, the standard in any school subject or form of pupil achievement may be a definitely formulated, although probably subjective, or even only a vaguely conceived, idea in the mind of the teacher or principal concerning his expectations of his pupils. In this sense, standards are extremely variable and differ from school to school, from teacher to teacher, and even, as his ideas and pupil groups change, from year to year for the same teacher.

The modern emphasis upon providing for each child as an individual the type of instruction best adapted to his abilities, interests, and present and future needs, rather than upon the molding of all pupils into the same achievement pattern, has reduced the reliance of school people upon standards. The attempt is rather to furnish maximum aid to each child in the development of his potentialities and to evaluate his achievement in terms of himself as an individual.

### **Establishing final validity and reliability**

The procedures discussed above, although sometimes complex and always time-consuming, are prerequisite to the final steps in the publication of a standardized test. After these steps have been carried out, the final forms of the test given to a representative group of pupils, and the norms derived on the basis of their scores, it remains for



the test maker to obtain final evidence concerning the validity and reliability of the test and the reliability of the norms. Although careful and accurate work on the preliminary steps should make reasonably certain that these important criteria will be satisfied, it is nevertheless essential that these steps be performed as a final check and that their results be reported to users and prospective users of the test to enable them better to evaluate it.

The interlocking and complex nature of some of these final steps makes necessary only a brief presentation here of the most important aspects of this final checkup. For the beginning student, Chapter 14 presents an adequately comprehensive treatment of the methods of determining the validity and reliability of tests.

*Validity of the test.* If the test is one for which validity coefficients of one or more of the types discussed in Chapter 4 will be meaningful, such validity coefficients should be obtained. In some instances this might require the administration of some other test to the group of pupils on which the test is standardized and the comparison of results obtained. In other cases it may require a comparison of test scores with course marks or teachers' ratings of pupils. Evidence concerning validity is also found in test norms that consistently show higher average scores with advancement in age and grade placement of the standardization group of pupils. In any event, evidence must be obtained directly or indirectly to show that the test measures what it purports to measure.

*Reliability of the test.* Test reliability must be established for the final form or forms of the instrument. Techniques such as those presented in Chapter 14 and even some more refined methods might be used. The reliability of measurement, which gives an indication of the accuracy of the scores obtained on the test, should also be determined. The purpose of such procedures is to establish the fact that the test measures accurately and consistently.

*Reliability of the norms.* One of the major problems in the derivation of norms for standardized tests has to do with the reliability of the norms themselves. Possibly the statement of this problem would be made clearer if the word *universality* were substituted for reliability in the foregoing statement. Reliability implies consistency, but universality reflects the generally representative nature of the norms. An otherwise excellently made test may be limited in its usefulness through the fact that the norms are not sufficiently repre-

sentative. It may be that it is hopeless to expect to produce norms which are so generalized that they represent suitable bases of comparisons wherever they may be found or used. However, the only hope lies in one of two directions. One is to sample so widely in the possible areas of population likely to use the test that practically every type and character of pupil and school situation is included. The other is to recognize the practical difficulties in the way of making a general norm fit all types of situations, and to select the population used in the derivation of the norms to represent deliberately chosen types of school situations. It would be impractical to expect pupils from small school systems with little or no laboratory or shop equipment to achieve at a level comparable to that expected of pupils from schools with large, well-equipped laboratories. The solution of this problem may lie in the establishment of representative norms for different types of courses and schools.

## 2 PRACTICAL USES OF STANDARDIZED TESTS

The value of the educational test is directly proportional to the extent to which the results from its use are translated into improved instructional, guidance, and administrative practices in the school. If these practices bring about improvement in the conditions under which teachers teach and children learn, the primary functions of school administration and supervision will have been realized. While the problems of securing these results in the most effective and economical manner are treated in this chapter primarily in terms of the test as an instructional device, the guidance, supervisory, and administrative uses of these instruments must not be slighted.

### Instructional uses of standardized tests

The value of standardized tests for guidance, supervisory, and administrative, and research purposes has been emphasized so generally that very often the classroom teacher overlooks their real value in the solution of his own instructional problems. Yet this is where the most vital and important uses of such tests are to be found. The development of reliable, valid, and highly detailed measuring instruments has caused the teacher to modify his previous conceptions of the uses of standardized tests. Earlier experience with the more



formal types of educational tests sometimes led the teacher to feel that tests were merely time-consuming devices used for checking up on his teaching efficiency, from which he received little or no constructive help in the improvement of his instruction. Quite in contrast with this idea, the more modern conception of standardized tests implies their continuous use as instruction progresses. This means a continuous testing program, for experience with the other conception of the use of tests indicates that only through continuous testing will standardized tests ever come to function at their highest efficiency as instructional instruments in the classroom.

*Class analysis and diagnosis.* Very often a teacher, at the beginning of a school term, wishes to obtain advance information concerning the proficiency of his classes in certain subjects and their general preparation for the work. It is essential for him to know their weaknesses and their strengths in some detail in order so to direct their work that the best results will be obtained. He needs to know the background his pupils have been given for the work they will be expected to master during the ensuing year. Most modern standardized tests permit this type of use. It is not always necessary for a teacher to employ a special diagnostic test to secure the required general data on the relative abilities of the class. For example, the ability to read silently is the basis of proficiency in so many subjects that the teacher should certainly secure a picture of the reading ability of the class. The results should indicate whether the class as a whole or the individual members of the class are able to interpret the printed page with facility, and so carry on their work without great assistance. It is also possible to test in a like manner for other general qualities, as well as for mastery of specific knowledge outcomes.

Not only is this preliminary general diagnosis of great value to the teacher, but it has also been found desirable and valuable to check progress or advancement from time to time by means of objective tests. Tests of achievement will reveal whether the class as a group is moving together, or whether there are more or less well-defined sub-groups that seem to need special attention. Frequently such conditions furnish a justifiable basis for dividing a grade or class into sections for such corrective treatment. Where classes are divided into several sections, as is often done in larger schools, many competent educators feel that the pupils should be arranged so that groups of approximately equal ability are placed

together. The objective test is the best means for making this adjustment so that pupils can move forward in groups of nearly equal proficiency.

An illustration of this need is given in Table 6, which shows the rather startling range of ability found in a typical ninth-grade class for results from the *Terman-McNemar Test of Mental Ability*. A teacher confronted with a class ranging in mental ability from above twelfth grade to below seventh grade is faced by a hopeless task if he attempts to bring all members of this class up to the same level of proficiency. Particularly is this true when the range of ability is wholly unsuspected or measured, as it is in so many cases, by guess rather than by reliable tests. The systematic use of tests for the purpose of identifying class needs constitutes a real source of professional protection to the classroom teacher.

TABLE 6. Distribution of mental ability in a ninth-grade class in terms of average grade placement

Grade Location	Number of Pupils
Above 12	6
12	4
11	12
10	8
9	19
8	8
7	6
Below 7	4
Total Number of Cases	67

*Individual pupil diagnosis.* Closely connected with the use of the test for pupil guidance is its use for the determination of the difficulties and variabilities of each individual pupil. While in general individual differences may not be so marked as to preclude reasonably efficient class instruction, the more that is known about each child's weaknesses and strengths the greater are the possibilities for success on the part of the teacher instructing the group. The test results should be studied especially in the light of each pupil's individual attainments and points of difficulty. The critical analysis of each pupil's test scores may very likely be a means of clearing up



wholly unsuspected troubles that would otherwise continue to hamper the child and to reduce his chances for proper advancement. Although this type of individual analysis has great possibilities, it becomes increasingly valuable when it is definitely tied up with remedial material so devised that each child may be aided in correcting his own weaknesses.

Examples of two diagnostic profile charts are given in the accompanying illustrations of interpretative materials provided with two achievement tests of the survey type. Both profile charts serve analytic or general diagnostic rather than specific diagnostic functions, but it should be remembered that survey tests can be diagnostic only in the broad sense discussed in Chapter 3. The profile chart for the *California Reading Test* furnishes places for recording graphically evidence of a pupil's grade placement on total reading, on the two sub-total measures of vocabulary and comprehension, and on seven part scores. The chart for the *Gray-Votaw-Rogers General Achievement Test* furnishes positions for the graphic recording of the pupil's grade placement on total achievement and on achievement in ten separate major areas of elementary-school achievement.

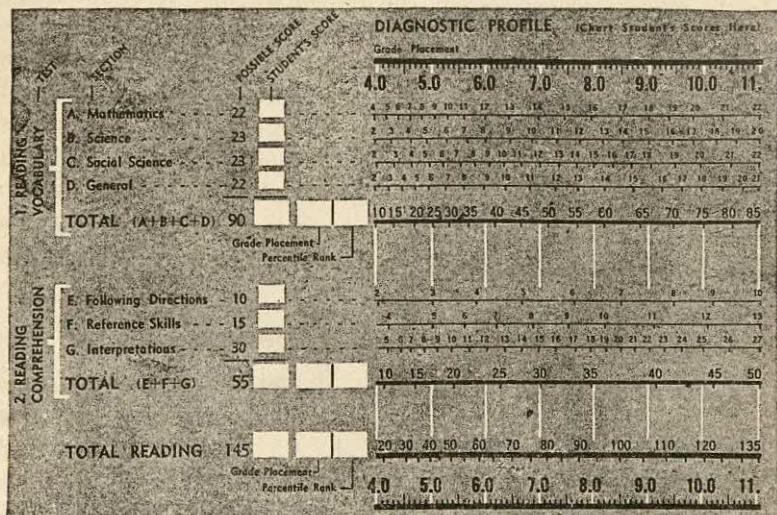


Fig. 3. Diagnostic profile chart for the California Reading Test<sup>5</sup>

<sup>5</sup> Ernest W. Tiegs and Willis W. Clark, *California Reading Tests*, Intermediate. California Test Bureau, Los Angeles, 1950.

## Guidance uses of educational tests

Schools are under constant criticism for their apparent failure to identify the special abilities of their pupils and to challenge these children to greater efforts. This is one aspect of educational guidance. Furthermore, it is charged that little or no attempt is made to direct children away from fields in which they apparently have little aptitude. With the modern objective devices now available for the measurement of general as well as specific abilities of children, neither of these situations needs to exist. Teachers, principals, and administrators have found that test records obtained early in the institution's contact with the pupil prove to be extremely valuable aids in handling disciplinary cases and in helping pupils to adjust themselves in many other ways. Most disciplinary problems arise through the failure of the school system properly to stimulate and occupy the pupil's mind. Many of the reasons for such difficulties may be made clear to the teacher by the wise use of properly selected tests. The necessary adjustments can then be made to correct a situation that need not exist if properly handled.

In the modern school system, the guidance center serves also as a testing bureau. It must work in closest coordination with the administrative office in the maintenance of an efficient child accounting system. Among the most important items entered in the records will be all types of readily interpreted results from physical and mental tests, aptitude and prognostic tests, personality inventories, and survey, analytic, and diagnostic achievement tests. The guidance specialist must therefore necessarily be well trained in the use and interpretation of measurement and evaluative tools and techniques.

## Administrative uses of standardized tests

During the early period of the growth and development of educational tests school administrators were the ones most directly concerned with their possibilities. Schools and classes were increasing in size, curricular offerings were expanding, educational costs were advancing, public interest in educational efficiency was increasing, and parents and teachers were growing more and more critical of the methods of evaluating pupil accomplishment and the marking systems then in use. Administrators themselves were becoming increasingly critical of some of their own practices. Students of educa-



This Child's		Elem. Sci. 1	Language 2	Literature 3	Spelling 4	Reading		Soc. Stu. 7	Health & Safety 8	Arithmetic		Total Aver.	This Child's	
Chro. Age	Educ. Age					Vocab. 5	Comp. 6			Reas. 9	Compu. 10		Educ. Grade	Sch. Grade
		95										95		
		90										90		
17-2													12.0	
16-10													11.6	
16-6													11.4	
16-2													11.1	
15-11													10.7	
15-8													10.5	
15-4													10.2	
15-0													9.9	
14-9													9.6	
14-6													9.4	
14-3													9.1	
14-1													8.8	
13-10													8.6	
13-7													8.4	
13-3													8.2	
13-2													8.0	
13-0													7.8	
12-9													7.6	
12-7													7.4	
12-4													7.2	
12-2													7.0	
12-0													6.8	
11-10													6.7	
11-8													6.5	
11-6													6.3	
11-4													6.2	
11-2													6.0	
11-1													5.8	
10-11													5.7	
10-10													5.6	
10-9													5.4	
10-7													5.3	
10-6													5.2	
10-5													5.1	
10-3													5.0	
10-2													4.9	
10-1													4.8	
10-0													4.7	
9-11													4.6	
9-10													4.5	
9-9													4.5	
9-7													4.4	
9-6													4.3	
9-4													4.2	
9-3													4.1	
9-2													4.0	
9-1													3.9	
9-0													3.8	
8-11													3.8	
8-10													3.7	
8-9													3.6	
8-8													3.5	
8-8													3.5	
8-7													3.4	
8-6													3.3	
8-5													3.2	
8-4													3.1	
8-3													3.0	
8-2													2.9	
8-1													2.8	
8-0													2.7	
8-0													2.7	
7-11													2.7	
7-10													2.7	

Fig. 4. Individual educational chart for the Gray-Votaw-Rogers General Achievement Test <sup>6</sup>

<sup>6</sup> Hob Gray, David F. Votaw, and J. Lloyd Rogers, *General Achievement Tests*, Intermediate. Steck Co., Austin, Texas, 1950.

tion and taxpayers asked searching and frequently embarrassing questions. Communities demanded school surveys as a means of answering their own questions concerning the efficiency of their schools. Naturally during this period the administrator turned to the test as one objective means of handling his problems of public relations.

Modern school administration demands that the most objective and reliable evaluative instruments available be used to provide the answers to the problems continually arising in connection with the operation of a school system. Often it is impossible to separate the administrative use of a test from an instructional or a supervisory use. The standardized test may be required to establish the adequacy of a given system of assigning and reporting teachers' marks. A specific unit of instructional material or a new and unproved method of teaching may require experimental evaluation. The efficiency of the school system in terms of pupil growth per unit cost may require demonstration. These and other possible administrative uses of standardized tests by the administrator are listed in Section 3 of this chapter on planning the testing program. Only three administrative uses are discussed here.

*Pupil gradation and placement.* Administrators, supervisors, and teachers find the problem of pupil placement one of the most difficult situations they have to face. The indefinite lines of division between the grades and the wide overlapping of ability between grades reveal that the typical techniques of pupil classification now in use are extremely crude. This could scarcely be otherwise in view of the methods commonly used. The proper grade placement of pupils implies that insofar as possible individuals who are normal for their group should be placed together. This means that pupils who are approximately alike in their chronological age, their educational achievement, and their physiological, mental, social, and moral development should, where possible, be placed together for instructional purposes. Not all of these qualities lend themselves readily to objective measurement, but a number of them do, and within these limits the results of objective measurements should be used in determining the pupil's placement in his group. Through the development of reliable grade and age norms, based upon the achievement of groups of children on standardized tests, a valuable instrument for the establishment of grade lines and for within-class grouping is made available.



By means of a simple procedure, results from several different tests of a battery can be combined into a graphic record and used as a valuable aid in pupil classification. An example of this procedure is given in Figure 4 on page 110. The technique is simple and valuable, regardless of whether a complete reclassification of the school or grade is planned, or merely the proper placement of a few new pupils entering the system for the first time.

*Group comparisons.* Since the earliest beginnings of group instruction, classroom teachers have wished to know just how their pupils have compared in attainment with other similar pupils and classes. Until standard tests were developed, it was practically impossible to secure this information. Now the giving of standardized tests in arithmetic, spelling, reading, or other school subjects makes fairly easy a comparison of the results from a class with the norms established for the subject and grade.

Comparisons with other classes within the system in which the teacher is working, within the same building, and even between different sections of classes in charge of the same or different teachers can be made on a basis of objective norms that have been derived for the various tests. Another comparison that is even more useful is that between the attainment of a class at the beginning and the end of a semester's or a year's work, or at shorter intervals in the course of a semester. Each of these comparisons has its own peculiar value in assisting the teacher to determine the relative attainment and progress of his class at a given time.

*Measuring the efficiency of learning.* Such general comparisons as are cited above are of great value in themselves, but equally important is the determination of ways and means by which the act of teaching itself may be improved. Ambitious teachers everywhere are looking for the best methods of instruction in their fields. Teaching methods, which in the last analysis should be studied in the classroom by the classroom teacher, can be evaluated effectively by means of standardized tests. Instructional units within the course of study should also be evaluated. The measurement of the effect of certain types of drill exercises and the determination of the specific strengths or weaknesses of groups or classes illustrate the uncounted opportunities for the administrative use of these valuable instructional devices.

## Meaning and importance of educational diagnosis

Educational diagnosis implies the use of more or less technical procedures designed to locate specific learning and instructional difficulties, and if possible to determine their causes. For the medical expert, diagnosis means the careful and extensive observation of the patient under controlled conditions. It includes the use of professional instruments, such as the clinical thermometer, the stethoscope, and the microscope, which make possible exact and objective observations. It means the assembly of a complete case history of the background of the difficulty leading up to the present physical crisis. It is based on the examination and analysis of many similar cases, in order that common factors may be identified. For the teacher, diagnosis has many of the same implications, but unfortunately much of the exactness, objectivity, and precision of the medical diagnostician's instruments appear to be missing in the teacher's equipment. Even today only a few objective measuring instruments capable of rendering reasonably precise diagnosis are available to the pedagogical diagnostician. The well-prepared modern teacher now has at hand reasonably adequate statistical techniques; analytical diagnostic tests in different subjects; diagnostic charts; instruments and devices for measuring aural acuity, visual acuity, eyedness, muscular imbalance in the eyes, binocular vision, and binocular fusion; and many other highly important qualities that may account for a pupil's lack of progress in many fields of learning. It is thus apparent that diagnosis in education is moving rapidly in the direction of scientific accuracy.

The diagnosis of difficulties underlying educational accomplishments undoubtedly constitutes the high point in the supervisory and instructional uses of educational tests. Deficiencies of a general nature are revealed and brought to light by general survey tests. Specific weaknesses, and to a certain extent causes of weaknesses, are identified by the use of properly selected diagnostic tests. Practically all of the more exact types of diagnostic procedures, such as the location of defects in speech, hearing, and vision, are dependent upon educational test results for their initial steps. These points will be discussed much more thoroughly in the later chapters on reading and language.

*Analysis as the basis of diagnosis.* The successful development of the many sets of habits that constitute the bulk of school learning



depends upon the care with which the underlying and basic skills of the subjects themselves are recognized and utilized in the teaching. If it can be shown that teaching a child to add consists not only in developing the habit of responding automatically and correctly to the 100 basic combinations but also involves higher levels of skill, such as knowledge of the higher decade addition facts, bridging of the tens, control of the attention span, and carrying from one column to the next, the teacher's task is made obvious and objective. Similarly, if it can be shown that silent reading comprehension is not a single isolated ability but a composite of many elements, such as knowledge of word meanings, ability to get meaning from sentences, ability to arrange thought units and sentence units into logically organized wholes, and ability to find desired material quickly, the teacher has a real basis for his instructional procedures. Language is another basic subject in which many delicately balanced skills are interwoven in an extremely complex manner. Here again the elements of achievement in the total process must be identified. Blind trust in general practice on the total skill must necessarily give way to the exact identification and discovery of the particular points of pupil weakness as a basis for special emphasis.

Good diagnosis must parallel the processes of good teaching. Effective diagnostic materials in any school subject can be prepared only after the skills contributing to success in this field have been isolated and identified. Psychologically, the reason for this is that on the whole the child learns to do what he practices and not something else. Remedial work, accordingly, can function only when the point at which pupil mastery breaks down has been located. Thus the analysis must be penetrating and the diagnosis must be precise.

*Specific nature of diagnosis.* Diagnosis must be more exact than broad statements of general functions. It is not enough to discover that a child is unable to read silently. The exact nature of his handicap must be revealed before it is possible to undertake a remedial program. The more specific the diagnostic information revealed, the more exactly the remedial material can be made to fit the need. To return to a frequently used illustration, it is found by diagnosis that the child is unable to add, but unless the exact point at which his mastery of addition breaks down can be determined by the diagnosis teaching or remedial efforts are largely wasted. One of the outstanding reasons why more effective teaching and remedial work has not been done in certain fields is that no adequate analysis of basic skills

can be made or has been made. Concrete illustrations of this need are given in connection with a related discussion in this chapter.

*Importance of the diagnostic use of test results.* Tests as such are incapable of improving instruction because of any inherent power. Existing conditions are merely revealed by them, and these with the limitations implied by the validity and reliability of the particular instruments used. Remedial or corrective teaching is the result of deliberate constructive effort by the teacher after the particular points of weakness in the instruction of the pupils have been revealed by the tests. The ease, clearness, and directness with which these needs are revealed by the tests are a measure of their real educational value. Too few existing tests are so constructed as to permit the interpretation of their results directly in terms of an effective remedial procedure. However, this seems to be no good reason for the failure of teachers to apply more directly the results of this work in testing to the improvement of their teaching practice. Just as the data revealed by the navigator's instruments require calculation and interpretation, so is it necessary to analyze test data carefully in order to make them the basis of a genuine remedial program.

The interpretation of test scores and the planning of remedial procedures are the most difficult parts of the use of standardized educational test results. Moreover, they are by far the most important parts. One of the greatest needs in education today is the provision for genuine diagnostic testing in all instructional fields, supplemented by valid remedial work designed to correct the weaknesses and defects of individual pupils as revealed by the tests. It is important to learn, as a result of using tests in the classroom, that a pupil or the entire class is below the norm in the subject, but unless it is learned with some exactness what causes the low level of achievement the testing program will do little if anything more than supply interesting information. Teachers and supervisors have a right to expect that something more constructive will be provided in exchange for the time required for classroom testing.

*Diagnosis the basis for remedial work.* Accurate diagnosis of class and individual pupil difficulties, coupled with application of specific remedy, is the heart of enlightened use of exact methods of teaching. The success of the remedial or corrective teaching depends upon the accuracy and detail with which the specific skills involved in successful achievement in the subject are identified and isolated in the test. Tests of the general survey type, or tests that report single unanalyzed



scores, cannot supply this information in sufficient detail. Specific examples of these points will be found in later chapters dealing with special subject tests.

*Diagnosis as the basis for preventive work.* An examination of the number and types of skills identified as a result of the diagnostic methods discussed in the preceding section leads to a suggestion of a still more constructive use of analytic and diagnostic test results. Diagnosis as applied in education has taken on a meaning indicative of a breakdown in method, a failure of instructional techniques to function. Unquestionably, one of the basic purposes of diagnosis is the location of weaknesses and the determination of their causes, but there is nothing in the method that precludes its use in the prevention of weaknesses through anticipation of their causes. Out of the knowledge gained through the use of diagnostic procedures should come the basis for preventive work of all types. It is quite noticeable that the major emphasis in the fields of dentistry and medicine is not on correction but on prevention. The existence of a weakness implies a failure at some point in the program. The discovery of it should not be marked as important merely because it is then possible to correct it. The real importance in the discovery should lie rather in the prevention of its reappearance elsewhere under similar conditions.

Another illustration from the field of medicine may make this point somewhat more concrete. In every medical examination for diagnostic purposes, a complete analysis is made and an exact case record of all observations is kept. Out of the analysis of these records has come a better understanding of the causes and characteristics of certain types of human ailments. Out of this same type of analysis has also come the basis for much of the preventive work that characterizes modern medical science. In a similar way, accurate and detailed educational diagnosis may ultimately offer the basis for the development of a program of preventive work in education. For example, if, after diagnosing the addition of fractions in the fifth grade, it is found that the failure of pupils to reduce fractions in the answers is a common weakness, the obvious thing to do is to correct the defects at once and then proceed to reconstruct the first instruction so that the following year the causes for this particular weakness may not operate so powerfully. Similarly, any weakness identified now should afford the basis for decisions calculated to reduce the probability of their recurrence in the future.

## The place of remedial instruction

*General practice exercises vs. remedial drill.* There are in general two ways of maintaining a high level of pupil achievement in any subject after direct instruction has been discontinued. These are (1) broad, general drill with no integral units of testing to discover breakdowns in pupil mastery, and (2) systematic remedial drill devices to fight forgetting, plus diagnostic testing to discover the exact causes of weaknesses when such weaknesses begin to cause poor work on review drills. The first method involves the systematic use of properly distributed general practice over the complete function. The second involves the periodic location of the specific defects of each pupil by means of diagnostic tests and the immediate correction of these defects by the use of properly constructed remedial drill.

Unquestionably the latter is the more economical method of maintaining mastery of desired skills on the part of a pupil. It is obvious that general review is valuable at times, but just to review with no specific idea of what the review is to accomplish is too naïve and hopeful to be effective. The program that coincides most closely with the experience of successful teachers and with a sound psychology of learning calls for the following steps in approximately the order indicated: (1) *teach*, (2) *review*, (3) *test for weaknesses* whenever they appear, and (4) *follow with remedial drill units* on the specific weaknesses revealed by the tests. It may be worth while to note that material so constructed as to be effective for remedial purposes is also sound to use for initial instruction. In fact, the chief distinction between good content for initial teaching purposes and remedial drill purposes lies in *when* they are to be used. The most effective remedial drill for the pupil who does not have an adequate sight-meaning vocabulary for silent reading purposes is drill on the vocabulary he should have learned in the first place.

*Necessity for valid drill for each identified skill.* If remedial work is to be effective, drills of established validity must be provided for each specific skill which conditions achievement in the subject. The validity of drill material depends to a large degree upon the accuracy and completeness with which the analysis of skills is made. Difficulties in subject units which can be identified in only a vague manner cannot be remedied except by chance. Drills must closely parallel the skills they are supposed to remedy. If mastery of a certain minimal vocabulary is essential to effective silent reading comprehension, then



drill on those particular words that constitute special weakness should take precedence over other drill.

Theoretically, perfect validity of drill material can be achieved only by taking a 100 per cent sampling of all of the possible basic facts or skills in the particular field. Naturally this is impossible in certain cases, but it is nevertheless often possible to take such a large sampling that all of the most frequently used and most important facts are included. Subject fields vary widely in the ways in which they lend themselves to sampling of this kind. In fields such as reading or language, a complete sampling is almost impossible to obtain. On the other hand, many of the basic facts in arithmetic are so readily identified that they may be sampled 100 per cent without difficulty.

Properly designed remedial and corrective drill material wastes no time on skills that need no practice, but strikes directly at the heart of the trouble. Remedial drills in which careful control is kept over the distribution of practice on the basic skills are almost certain to be more effective than random exercises, even assuming in both cases that suitable motivation for improvement is provided. That drill will be most productive which most nearly provides a complete coverage of the skills of basic importance in the hierarchy of habits upon which successful achievement in the subject depends. Poorly organized drills may or may not deal with all possible weaknesses, but they are almost certain to waste time on skills that are not in need of drill. The validity of the drill depends upon the degree to which this sampling covers the basic or fundamental skills and the degree to which the exercises themselves actually develop the skills they purport to develop. There are a number of places in which this complex chain may break. The task of diagnostic and remedial treatment is to locate and repair quickly those links of the chain that have snapped under stress, or have rusted out through lack of use. Correctly designed remedial material will not only parallel valid drill on the correct skills, but it will also cover all of the basic aspects of the skill. Furthermore, it should acquaint the child with the most important variants of each situation.

Effective remedial material must not only cover in a valid manner all of the basic or underlying skills upon which achievement in the field depends, but it must provide a means for bringing about a gradual union of these component elements into the total function. It is

entirely possible that a mastery of the subsidiary skills involved might result in only a partial control of the end product, if that goal were not reached by the gradual bringing together of each distinct skill in its relation to the whole process.

### 3 PLANNING TESTING PROGRAMS

#### Steps in a testing program

The following brief outline is presented here as a suggestion to the teacher, supervisor, or administrator for the general organization of the testing program:

1. Select and state a clear-cut teaching problem in the solution of which test results appear to be essential.
2. Secure the cooperation of the school staff in the attack on the problem.
3. Determine what types of test data will be valuable in the solution of the problem.
4. Select the best available tests for the purpose.
5. Make careful preparation and then administer the tests.
6. Score the tests as quickly, accurately, and economically as possible.
7. Tabulate the scores and analyze and interpret the results.
8. Use the results and the interpretations in the elimination or improvement of the conditions revealed, depending upon the nature of the problem.

*Clear-cut teaching problem as basis.* One of the most common errors made by teachers and supervisors is the inauguration of a testing program without first formulating a clear-cut problem the solution of which can be most advantageously reached through the use of tests. The problem should be sharply defined, for the testing program will thus be more limited in extent and more intensive. The best constructive supervisory work will result from careful and intensive cultivation of a limited field. If the work is undertaken in this way, much time will be saved and one of the most common criticisms—that the time of the pupil and teacher is taken for the testing and nothing ever comes of it—will be avoided. Both teacher and pupils have a right to profit from a knowledge of the conditions revealed by the testing.

*Illustrations of problems.* Problems suitable to form the basis for a testing program are to be found in almost all the fields of educa-



tion. Frequently these problems overlap. It is not uncommon for one test or a series of closely allied tests to contribute to the solution of several problems. The problems listed below are classified in accordance with their interest to teachers, administrators, and supervisors. It is clear, of course, that the list is not exhaustive.

#### A. PROBLEMS PRIMARILY OF INTEREST TO TEACHERS AND SUPERVISORS

1. The discovery and diagnosis of defects of individual pupils in the various subjects or in particular phases of a subject as the basis for a remedial program.
2. The determination of how the achievement of the pupils and the class compares with the norms in the different subjects.
3. The determination of the progress or growth of the class in the different school subjects over a given period.
4. The determination of whether different phases of subjects are being properly or unduly stressed, as indicated by relative accomplishments of the pupils.
5. The determination of the extent to which the pupils are working at maximum capacity.
6. The evaluation of the effectiveness of a given organization of instructional material.
7. The evaluation of the efficiency of a given method of instruction.
8. The experimental evaluation of textbooks.

#### B. PROBLEMS PRIMARILY OF INTEREST TO ADMINISTRATIVE AND SUPERVISORY OFFICERS

1. The determination of the misplacement of pupils in grades or sections.
2. The proper classification of new pupils entering the school system.
3. The division of classes into sections according to ability.
4. The selection of pupils for special classes, such as classes for exceptionally bright or exceptionally dull pupils or for pupils having special defects in certain subjects.
5. The determination of the efficiency of the school as a whole by comparison of scores with norms and with scores made by other schools or grades.
6. The determination of whether the proper emphasis is given to all subjects or whether some subjects are overstressed.
7. The comparison of different methods of instruction or comparison of new methods with the ones already in use.
8. The determination of the general achievement level of a grade, a school, or a system.

9. The measurement of the progress of a grade, a school, or a system for a semester, a year, or any given period.
10. The determination of whether or not a grade, a school, or a system is achieving what can fairly be expected in terms of current educational costs.
11. The evaluation of the effect of a special supervisory drive.
12. The compilation and use of educational and vocational guidance information.
13. The provision of answers to current local queries concerning the over-all efficiency of the school system.

*When to give tests.* The type of testing program followed depends somewhat upon the purposes the tests are to serve and the nature of the tests selected. If the tests used are survey tests of general achievement, they are usually given early in the school term and then perhaps again a few days before the end of the school term. This procedure permits the teacher to determine the improvement his pupils have made during this period. Tests that are used definitely for survey purposes and are given only once during the school year are usually administered at or near the end of the year. This is probably one of the least important times for tests to be given, since almost the entire school year is gone and there is little opportunity for the teacher to attempt to do anything about the conditions revealed by the tests. *If only a single cross-section of the school is taken, this should undoubtedly come early enough in the school year to permit the teacher to profit from the findings.* The periodic use of educational tests to measure class or individual pupil improvement is by far the most profitable practice.

A further refinement of the idea of giving tests early in the school year is found in their use immediately following the completion of instruction on a particular course unit. Unit achievement tests, each designed to measure a specific area of the course, are proving popular for this purpose with both teachers and pupils in certain subjects. By using these narrow-function tests immediately after the completion of the teaching of a specific instructional unit the teacher secures immediate information about the weaknesses of his class. Special inadequacies of instruction are thus made clear, and he can proceed at once to set up a remedial program before the class has moved on to other activities. This suggests the *continuous* use of tests as the basis for remedial work. The information provided by the use of these numerous narrow-function tests is also valuable in organizing future



instruction in such manner as to prevent the appearance of such weaknesses.

*Cooperative testing programs.* During the past fifteen years, cooperative testing programs have developed in many cities and states for the purpose of providing a coordinated attack upon measurement and evaluation problems. These programs are very different in organization, sponsorship, and objectives, but they typically provide testing services of such a nature that the participation of most teachers is limited either to the administration of the tests and use of results or alone to the use of results.

1. *City testing bureaus.* Bureaus of testing and measurement in a number of the larger cities maintain staffs of measurement and research specialists whose primary functions are to carry on planned testing programs and also to conduct related research studies. Frequently the cooperation of teachers is obtained in the administration of tests, and the results are made available to them for use with their pupils. Programs are frequently planned in cycles of several years, and tests in line with the total program may be given annually, twice a year, or at more frequent intervals.

2. *State-wide testing programs.* Testing and related services are now available to the schools of approximately three-fourths of the states through some public educational agency in each state.<sup>7</sup> Several of the states have two such programs. Frequently the testing programs are based on cooperative construction, administration, and scoring of the tests and uniform methods of reporting results in comparable form. In other cases available standardized tests are cooperatively administered and scored and the results are reported in as uniform a manner as possible. State-wide norms are frequently provided. New forms of tests are constructed or provided annually in some state-wide setups.

These programs and services vary widely among the different states, and include various patterns of achievement, intelligence, and personality tests. Some of the programs are conducted as scholarship contests, some are cooperatively sponsored by collegiate institutions that make use of scholastic aptitude test results of high-school seniors, some are conducted primarily for purposes of supervision, and still others are administered purely as services to the schools. Schools

<sup>7</sup> David Segel, *State Testing and Evaluation Programs*, U. S. Office of Education, Circular No. 320. Federal Security Agency, Washington, D. C., 1951.

in some states participate in the programs on a cost basis, and participation is most often optional for each school.

3. *Nation-wide testing programs.* Cooperative testing on a nation-wide basis is offered by various educational foundations, cooperative services, and commercial agencies for a wide variety of educational and mental tests. The services are sometimes provided primarily for a particular group of schools and in other cases are provided for any school wishing to obtain them. Reports are furnished to participating schools, and norms are often prepared on regional and nation-wide samplings of pupil test results.

#### 4 SELECTING TESTS

Test selection depends upon the type of testing program planned, for tests should be chosen that not only are within the proper subject field and at the appropriate level of advancement for the pupils but that also will serve the desired function. It should be pointed out that not all tests are appropriately named, however, and that too much dependence can easily be placed upon a test title. Accordingly, the student and teacher should learn to utilize critical standards in the selection of testing instruments.

#### Need for care in selecting tests

As has been implied above, the mere fact that a test is standardized guarantees neither its validity for the type of use prescribed by its author and publisher nor its validity for the use to which a teacher wishes to put it. A valid test of achievement should consist of items that are in harmony with the accepted objectives of the subject in question. Yet, it has been found that in five tests of English usage from one-sixth to more than one-half of the usages scored as wrong in the different tests are acceptable in terms of the standards acceptable to the National Council of Teachers of English.<sup>8</sup>

The teacher or administrator selecting standardized tests has a right to expect that accurate information will be furnished him by the author and publisher concerning the validity, reliability, and other criteria of a good examination. Valuable sources of evidence about tests are the *Mental Measurements Yearbooks*, published in new

<sup>8</sup> Karl W. Dykema, "On the Validity of Standardized Tests of English Usage." *School and Society*, 50:766-68; December 9, 1939.



editions at frequent intervals.<sup>9</sup> These yearbooks contain carefully edited descriptions and critical reviews of tests by subject and test specialists with which to supplement information about a test furnished by the author and publisher.

### Test rating scales

In the discussion of criteria for tests in a previous chapter, no attempt was made to evaluate in a definite manner any of the items that appear to affect test quality. Numerous rating devices that weigh the various items roughly in order of importance are available.

#### Standardized Achievement Test Rating Scale

Criteria	Maximum Ratings	Test _____		Test _____	
		Ratings	Reasons	Ratings	Reasons
1. Validity	20				
2. Reliability	10				
a. Adequacy	10				
b. Objectivity	10				
3. Practicality	5				
a. Administrability	10				
b. Scorability	10				
c. Economy	5				
4. Comparability	15				
5. Utility	5				
Totals	100				

#### Summary statement of major reasons for preference

<sup>9</sup> Oscar K. Buros, editor, (1) *The Nineteen Thirty Eight Mental Measurements Yearbook*. Rutgers University Press, New Brunswick, N. J., 1938; (2) *The Nineteen Forty Mental Measurements Yearbook*. Mental Measurements Yearbook, Highland Park, N. J., 1941; (3) *The Third Mental Measurements Yearbook*. Rutgers University Press, New Brunswick, N. J., 1949; and (4) *The Fourth Mental Measurements Yearbook*. Gryphon Press, Highland Park, N. J., 1953.

The assignment of point values to the different features of the tests is, of course, largely a subjective procedure. It is obvious that two different individuals using rating scales could not be expected to agree closely on the scores assigned to a particular test. However, in spite of these limitations, such rating scales are of very real value to the inexperienced teacher or student because of the definite way in which attention is called to the quality features of a test.

The accompanying rating scale is suggested for use when two or more standardized achievement tests are being considered for use in a situation in which the purpose is well defined and the pupils to be tested have been decided upon in advance. If the tests are comparably rated by the same person, the resulting total scores should lead to the selection of the test that will best serve the purpose. The scale is organized in terms of the criteria of a good examination outlined in Chapter 4, and the weights assigned the various criteria are thought to represent their relative significance in total test validity.

A supplementary check list for use before the rating scale is filled out appears in another accompanying illustration. Provided with spaces for recording significant information about a test, the check list should be used separately for each test under consideration. Such sources of information as the publisher's catalog, the test manual and other testing materials, and even the *Mental Measurements Yearbooks* and other sources of critical test reviews might well be consulted in filling out such a check list.

### Standardized Achievement Test Check List

Title \_\_\_\_\_ Copyright \_\_\_\_\_

Author(s) \_\_\_\_\_ Publisher \_\_\_\_\_

1. Validity (Measures what it *attempts* to measure; Proper purpose and level) \_\_\_\_\_

Analysis of \_\_\_\_\_

Recommendations of \_\_\_\_\_

Accomplishment of \_\_\_\_\_

Rise in success by \_\_\_\_\_

Social utility \_\_\_\_\_

2. Reliability (Measures what it *does* measure; Consistency)

Reliability coefficient(s) of \_\_\_\_\_ type, of size:

\_\_\_\_\_ on \_\_\_\_\_ cases in Grade(s) \_\_\_\_\_ from \_\_\_\_\_ schools in \_\_\_\_\_ states.

\_\_\_\_\_ on \_\_\_\_\_ cases in Grade(s) \_\_\_\_\_ from \_\_\_\_\_ schools in \_\_\_\_\_ states.

- a. Adequacy (Wide sampling of items in outcome(s) measured)

No. of booklet pages \_\_\_\_\_; No. of items \_\_\_\_\_; Testing time \_\_\_\_\_.

Types of outcomes measured \_\_\_\_\_



- b. Objectivity (Absence of subjectivity or bias in correct answers)

Types and numbers of items:

Recognition: Alternate-response \_\_\_\_\_; Multiple-choice \_\_\_\_\_;

Matching \_\_\_\_\_.

Recall: Simple recall \_\_\_\_\_; Completion \_\_\_\_\_.

Miscellaneous \_\_\_\_\_

3. Practicality (Practical considerations; Feasibility)

- a. Administrability (Ease of administering)

Working time on Part(s): I \_\_\_\_\_; II \_\_\_\_\_; III \_\_\_\_\_; IV \_\_\_\_\_;  
V \_\_\_\_\_; VI \_\_\_\_\_.

Directions \_\_\_\_\_; Preparation for giving \_\_\_\_\_; Pupil instructions \_\_\_\_\_.

Materials needed: Booklets \_\_\_\_\_; Answer sheets \_\_\_\_\_; Special pencils \_\_\_\_\_; Manual of directions \_\_\_\_\_; Other \_\_\_\_\_.

Administered by: Teacher \_\_\_\_\_; Specialist \_\_\_\_\_; Psychologist \_\_\_\_\_.

- b. Scorability (Ease of scoring)

Scoring key \_\_\_\_\_; Scoring directions \_\_\_\_\_; No. of separate scores \_\_\_\_\_.

Scored by: Clerk \_\_\_\_\_; Teacher \_\_\_\_\_; Psychologist \_\_\_\_\_; Machine \_\_\_\_\_.

- c. Economy (Cost in money and time)

Booklets reusable \_\_\_\_\_; Separate answer sheets \_\_\_\_\_.

Cost: Booklet \_\_\_\_\_; Answer sheet \_\_\_\_\_; Special materials \_\_\_\_\_.

4. Comparability (Bases for interpreting results)

Norms: Age \_\_\_\_\_; Grade \_\_\_\_\_; Percentile \_\_\_\_\_; Other \_\_\_\_\_.

Raw or derived scores \_\_\_\_\_; No. of duplicate forms \_\_\_\_\_.

Norms based on \_\_\_\_\_ cases.

5. Utility (Use to be made of results)

Need to be served \_\_\_\_\_

To be used in Grade(s) \_\_\_\_\_ or with \_\_\_\_\_

Planned use of results \_\_\_\_\_

Class record \_\_\_\_\_; Pupil profiles \_\_\_\_\_; Diagnostic aids \_\_\_\_\_; Instructional aids \_\_\_\_\_; Other special materials \_\_\_\_\_

## 5 ADMINISTERING TESTS

The general procedures for administering tests suggested below are common to most tests now in use. They are not intended to take the place of the directions accompanying the various tests that may be used. The directions for giving and for scoring supplied in the examiner's manuals that accompany the better tests should be rigorously followed in order to guarantee that the tests are given under standard conditions.

## Preparation for testing

Any individual who is reasonably skillful in discipline and who will carefully follow the directions accompanying the tests should be able to administer a modern educational test. Unless the test directions are extremely familiar, the examiner should study the manual carefully before attempting to give the test. If possible he should administer the test to some other person in order to gain further familiarity with the procedure. If this is not possible, the directions should be read aloud several times so that they may be followed easily as the test is given. Familiarity with the directions is essential if the standard conditions for the test are to be maintained, and valid comparison of results with the norms thus be made possible.

Pupils may be tested in ordinary classroom groups or in larger groups. If several grades are to be given the same test, time may be saved by moving all pupils into a larger room, care being taken that the seats and the desks are suitable.

Before the test folders are given out, the desks should be cleared and each pupil should be provided with a sharpened pencil, or, if the test is to be scored by machine, with a suitable electrographic pencil. A number of extra pencils should be available for emergencies during the examination. The room should be quiet throughout the test. No questions should be allowed during the test. A manner that is agreeable but that at the same time suggests authority should be cultivated. Pupils should be made to feel "at home" in taking the test. Pupils will look forward to taking tests without fear or nervousness if the tests are properly given and if no misconceptions about the meaning and use of the results are allowed to arise.

## Administration of the tests

Throughout the examination, directions should be given in a forceful manner and should be spoken slowly and with careful attention to emphasis. The voice should be just loud enough to carry to all parts of the room. The directions accompanying the tests should be followed *verbatim*. As far as possible disturbances within or without the room that might interfere with the administering of the tests should be prevented. To avoid interruptions, the teacher may prepare a card carrying these words: *Testing Going On, Please Do Not Disturb*. If



this card is hung on the outside of the classroom door, interruptions will be less frequent.

The time limits as set in the directions for giving the tests should be strictly observed. Tests should be timed to the second or the results may not be comparable to those others get when the exact time is taken for the test. In timing the test a stop watch is very desirable. If an ordinary watch is used, one having a second hand, so that the minute and second hands can be synchronized, is preferable. The following illustrative procedure will serve quite well if a stop watch is not available:

	Hr.	Min.	Sec.
(a) Record time starting signal is given . . . . .	11	18	20
(b) Add to this the time required for the test	—	15	00
(c) The sum is the time to signal a stop . . . . .	11	33	20

### Teacher responsibility

In the earlier stages of the development of standardized tests, it was believed that the most valuable results came from their use in a periodic survey by persons other than the classroom teacher. More recently it has come to be generally accepted that as many of the tests as possible should be given by the classroom teacher. This seems to be especially true in the case of tests that furnish information of special importance in the improvement of instruction. In addition to allowing the classroom teacher to become acquainted with the technique of testing, it gives him a first-hand opportunity to observe the reactions of his individual pupils in the various test situations. On this account, it is believed that, wherever test results are to be used definitely as a basis for the discovery of individual pupil difficulties, tests should as far as possible be administered by the teacher himself. However, where the test results are used for a survey of achievement in the entire school or system, it is less important for the teacher to have an intimate contact with the testing program. As a matter of fact, many school administrators prefer not to have the teachers give the tests when they are used for such survey purposes.

## 6 SCORING TESTS

The scoring methods and devices discussed below are those used more or less widely with various standardized tests. Other procedures which are not especially designed for specific standardized tests but

which are more widely used for the informal objective examination are discussed briefly in Chapter 7.

### Hand-scored tests

The scoring of most modern standardized tests is made almost wholly objective by the use of stencil-type keys that fit the test or the separate answer sheets. The answer keys and directions for scoring each specific test should be followed rigorously. Scores should be obtained in exactly the manner prescribed by the test authors, in order that they may be compared directly with the norms that have been derived for the tests. It is best that all calculations be performed twice, and that all transcribed records be checked against the pupils' test papers to make sure that no errors have been made.

Hand-scoring keys of several types are used, among the most common being strip keys, cutout stencils, and transparent stencils. When answers are given in column form, strip keys that have the correct answers spaced on narrow strips of cardboard to correspond in spacing with the items of the test may be placed alongside a pupil's work for rapid scoring. When answers are scattered over a page and whenever the answer itself is the only point requiring the attention of the scorer, stencils having correct answers adjacent to apertures cut so that they will fall directly over the pupil's answers as the key is placed over the test also permit rapid scoring. Transparent stencils are similar to the above type, but are inconvenient because they do not permit the scorer to check the pupils' answers directly on the test paper or answer sheet.

The matter of responsibility for scoring hand-scored tests constantly arises as an administrative problem in the smaller schools. Teachers are likely to feel that the responsibility for scoring standardized tests given for supervisory purposes should not fall to them. A part of this difficulty arises through a failure on the part of the administrators to make perfectly clear to the members of the teaching staff their responsibility toward this type of work at the beginning of their terms of service. Most teachers, if given a suitable amount of time, do not seriously object to scoring test papers for their classes, particularly when they come to realize that this work may reveal information that will be extremely significant to them in the improvement of their teaching practices. If a real interest in the outcome of the testing program is stimulated by the supervisory officers, there



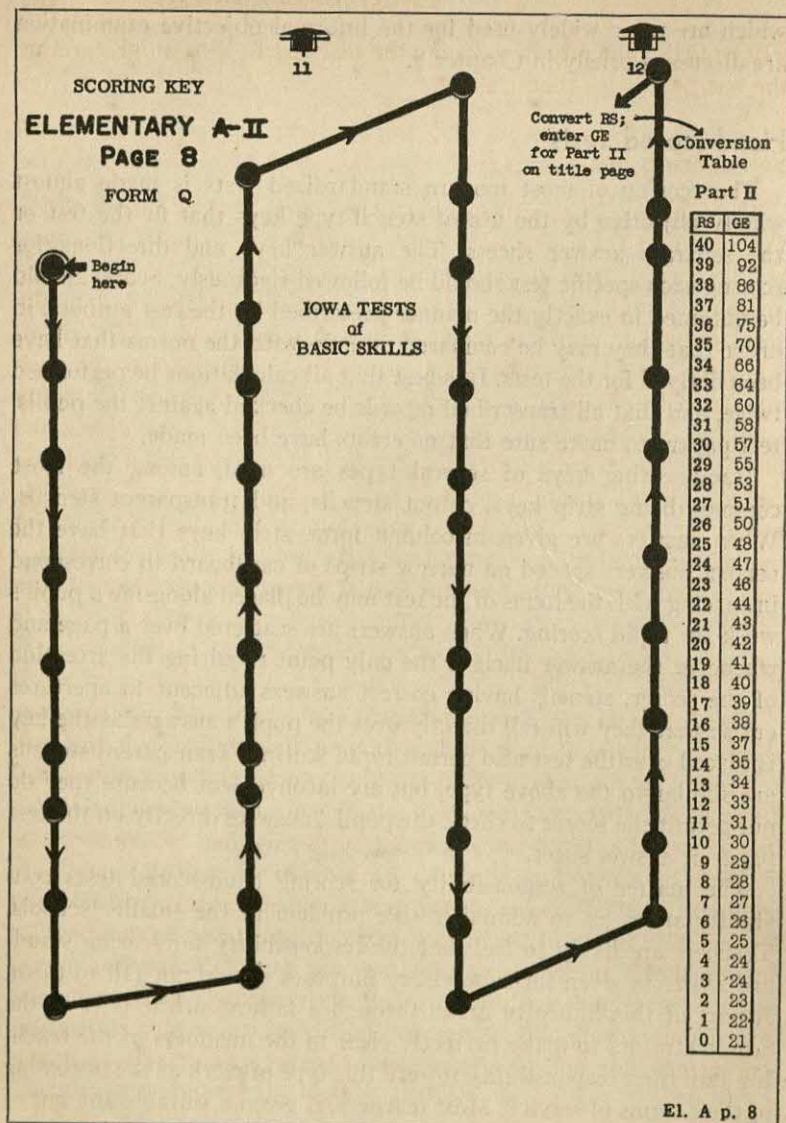


Fig. 5. Cutout scoring stencil for the Iowa Every-Pupil Tests of Basic Skills<sup>10</sup>

<sup>10</sup> *Iowa Every-Pupil Tests of Basic Skills, Elementary*. Houghton Mifflin Co., Boston, 1947.

will be little difficulty in inducing the teachers to help in interpreting the test papers for their classes.

### Self-scoring tests

The *Clapp-Young Self-Marking Tests*<sup>11</sup> and *Scoreze* answer sheets for certain of the *California Achievement Tests*<sup>12</sup> consist of answer booklets with carbon or wax transfer paper so placed that the pupil's answers to multiple-choice items are impressed on the back of the sheet on which he marks them. Each four-page booklet, kept closed while the pupil takes the test, is opened for scoring. Correct answers appear in the designated spaces on the back of the sheet for ready counting, while incorrect answers appear outside of the designated positions. The *Scoreze* forms also provide an original and a duplicate copy of the diagnostic profile as well as grade, age, and percentile norms for the test. The *Clapp-Young* folders are adapted for direct use with a number of standardized tests and are also available in a generalized form for use with informal objective examinations.

### Machine-scoring devices

The International Test Scoring Machine<sup>13</sup> scores pupil answer sheets by means of an electrical current flowing through the lead deposited by the pupil's electrographic pencil on the answer sheet. Items of the alternate-response, multiple-choice, matching, and modified completion types can be scored by this method.<sup>14</sup> Scores can be obtained by experienced machine operators at the rate of 700 or more per hour. Special answer sheets are provided and directly adapted for use with many of the newer standardized tests, while standard answer sheets in a variety of styles are available for the use of teachers or schools wishing to adapt their locally constructed tests to machine-scoring. The accompanying illustrations picture the test scoring machine and give examples of both types of answer sheets.

<sup>11</sup> Published by Houghton Mifflin Co.

<sup>12</sup> Published by California Test Bureau.

<sup>13</sup> Manufactured by International Business Machines Corporation, New York.

<sup>14</sup> *Methods of Adapting Tests for Machine Scoring*. International Business Machines Corporation, New York.



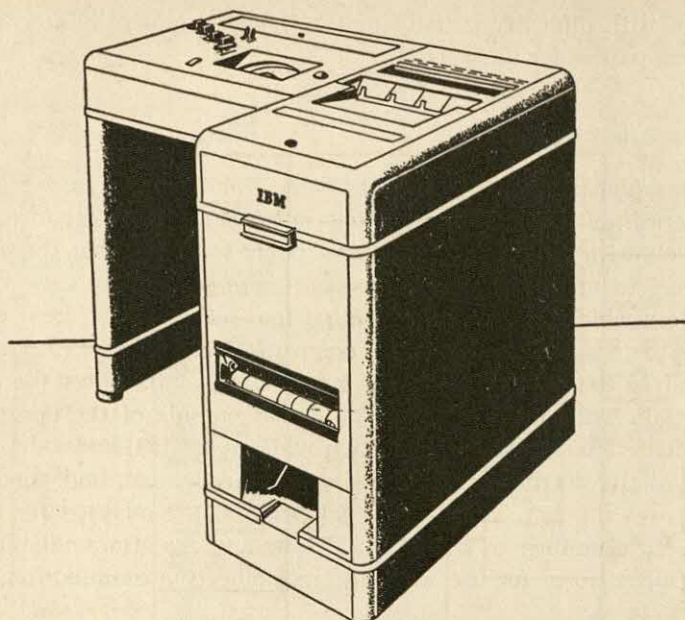


Fig. 6. International Test Scoring Machine

### Use of separate answer sheets

Prior to the development of machine-scoring devices, separate answer sheets for hand-scoring were used quite widely with teacher-made tests. The perfection of scoring machines in recent years has stimulated the use of separate answer sheets with many of the modern standardized tests. The need for long tests, required by demands for improved reliability of measurement with the resulting large and expensive test booklets, has done much to popularize the use of separate answer sheets for purposes of speed and economy, if for no other reasons. This is especially true in large school systems having access to mechanical scoring devices. Many different types of separate answer sheets have been developed which are adaptable to a wide variety of testing techniques. Such generalized or special answer sheets are also well adapted to hand-scoring with stencil keys of the cutout variety.

An exhaustive study of the effect on test validities and reliabilities of the use of separate answer sheets leads to the conclusion that the





separate answer sheets can justifiably be used.<sup>15</sup> Dunlap reported that there is no evidence to show that separate answer sheets cannot be used successfully with pupils in grades as low as the fourth. Some recent standardized tests provide separate answer sheets for pupils in the fourth and higher grades. However, for tests that are rather complex and require a complicated answer sheet, it is preferable that separate answer sheets should not be used much below the junior-high-school level.

## 7 ANALYZING AND INTERPRETING RESULTS OF TESTING

A complete discussion of the statistical techniques used in analyzing scores resulting from the administration of tests is given in Chapters 12 and 13. Accordingly, this problem will not be discussed here except to make the pertinent remark that the modern teacher is expected to understand and to be able to use such statistical techniques so that he will be able to obtain maximum values in using the results from tests given to his pupils.

The results of testing are interpreted by the use of norms and also by the use of certain derived scores that are dependent upon norms. A discussion of the derivation, and to a certain extent the application, of norms for standardized achievement tests appears in an earlier section of this chapter. Chapter 13 presents a rather complete discussion of derived scores and norms.

### Topics for Discussion

1. What are the distinctive features of the standardized test?
2. Show how the process of standardization involves much more than the mere establishment of norms for a test.
3. Indicate why the validation of content for standardized tests is more difficult for some school subjects than for others.
4. Show how discriminative power in a test item contributes to its validity.
5. What reasons can you suggest for the preparation of several equivalent and interchangeable forms of a standardized test?
6. Discuss the major types of test norms and illustrate each.
7. What factors appear to determine the type of norms that should be supplied with a standardized test?

<sup>15</sup> Jack W. Dunlap, "Problems Arising from the Use of a Separate Answer Sheet." *Journal of Psychology*, 10:3-48; July 1940.

8. What is the importance of determining the validity and reliability of standardized tests in their final forms?
9. What should be the teacher's responsibility toward the use of standardized tests in the classroom?
10. Suggest a procedure by which properly designed tests may be used for individual pupil diagnosis. For class diagnosis.
11. In your opinion, what types of tests (intelligence, aptitude, general achievement, diagnostic or analytic, personality) should the teacher be encouraged to use most freely? Why?
12. Why is it desirable to have a clear-cut problem in mind in initiating a testing program?
13. What is the possible contribution of a state-wide or other type of cooperative testing program to the solution of local testing problems?
14. What reasons can you advance for the failure to develop adequate diagnostic and remedial materials in all subject fields?
15. Select a school subject and show how the basic skills may be identified (diagnosed) in a way similar to that suggested in the discussion in this chapter.
16. In a school field that you are likely to teach (your major or an important minor), suggest a number of specific skills that enter into successful work and parallel this with suggestions for remedial treatment.

## Selected References

- BROOM, M. E. *Educational Measurements in the Elementary School*. New York: McGraw-Hill Book Co., Inc., 1939. Chapter 13.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949.
- BURT, CYRIL L. *Handbook of Tests for Use in Schools*. Revised edition. New York: Staples Press, 1948.
- CONRAD, HERBERT S. "Norms." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 795-802.



- CONRAD, HERBERT S. "The Experimental Tryout of Test Materials." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 8.
- COOK, WALTER W. "Achievement Tests." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 1461-78.
- COOK, WALTER W., AND OTHERS. "The Functions of Measurement in Education." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Part I.
- DAVIS, FREDERICK B. "Item Selection Techniques." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 9.
- FINDLEY, WARREN G., AND SMITH, ALLAN B. "Measurement of Educational Achievement in the Schools." *Review of Educational Research*, 20:63-75; February 1950.
- FLANAGAN, JOHN C. "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from the Data at the Tails of the Distribution." *Journal of Educational Psychology*, 30:674-80; December 1939.
- FROELICH, CLIFFORD P., AND DARLEY, JOHN G. *Studying Students: Guidance Methods of Individual Analysis*. Chicago: Science Research Associates, 1952. Chapter 11.
- HILDRETH, GERTRUDE H. *A Bibliography of Tests and Rating Scales*. Second edition. New York: Psychological Corporation, 1939.
- HILDRETH, GERTRUDE H. *A Bibliography of Tests and Rating Scales—1945 Supplement*. New York: Psychological Corporation, 1946.
- LEE, J. MURRAY. *A Guide to Measurement in Secondary Schools*. New York: D. Appleton-Century Co., Inc., 1936. Chapter 2.
- MCCONN, MAX. "The Uses and Abuses of Examinations." *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Co., 1936. Chapter 9.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. Chapter 13.
- ORLEANS, JACOB S. *Measurement in Education*. New York: Thomas Nelson and Sons, 1937. Chapters 6-8.
- REMMERS, H. H., AND GAGE, N. L. *Educational Measurement and Evaluation*. New York: Harper and Brothers, 1943. Chapter 10.
- ROSS, C. C. *Measurement in Today's Schools*. Second edition. New York: Prentice-Hall, Inc., 1947. Chapter 4.
- SEGEL, DAVID. *State Testing and Evaluation Programs*. U. S. Office of Education, Circular No. 320. Washington, D. C.: Federal Security Agency, 1951.

- TRAXLER, ARTHUR E. "Administering and Scoring the Objective Test." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 10.
- TRAXLER, ARTHUR E., AND OTHERS. *Introduction to Testing and the Use of Test Results in Public Schools*. New York: Harper and Brothers, 1953.
- VAUGHN, K. W. "Planning the Objective Test." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 6.
- WOOD, BEN D., AND HAEFNER, RALPH. *Measuring and Guiding Individual Growth*. New York: Silver Burdett Co., 1948. p. 274-82, 320-27.



## ***Constructing and Using Oral and Essay Tests***

THE METHODS of using the oral and essay examination and the characteristics of these types of subjective tests mentioned below are the basis for the discussion of this chapter:

- A. Extent and importance of classroom testing.
- B. Limitations and advantages of the oral quiz.
- C. Place of the oral quiz in the schools.
- D. Limitations of the essay examination.
- E. Advantages of the essay examination.
- F. Improving the essay examination.

The problems involved in the construction, selection, interpretation, and use of standardized tests have been discussed in the previous chapter. This chapter and the one that follows deal with the teacher-made or classroom test, as distinguished from the standardized test. From the point of view of many teachers, the classroom test constitutes the major problem of measurement.

### **1 CLASSROOM TESTING**

#### **Extent of classroom testing**

Every teacher is faced with constantly recurring problems of measurement and evaluation in the classroom. Not all such problems are best solved by objective tests, for evaluation techniques of rela-

tively subjective types also have their place in the classroom. However, each teacher spends days of time each year in preparing and scoring tests and in analyzing and interpreting the results. It has been estimated that teachers average more than a week of school time annually in work with tests.

### Need for improvement in classroom testing

The study of standardized tests thus far in this volume must make it apparent that by their very structure and use standardized tests do not meet all classroom needs for evaluation and measurement. In the first place, such tests do not equally well serve in all schools because of differences in emphasis and points of view resulting from the varying characteristics and educational needs of different communities. Also, classroom testing is sometimes important in an area so narrow or so specialized that no available standardized test fills the need. Again, teachers sometimes feel that standardized tests overstress factual knowledges and neglect what they believe to be important—the ability to organize and apply facts. For these reasons, written examinations prepared by local teachers, or at least within the local school system, will undoubtedly always be needed to meet the demands for complete and valid measurement of educational achievement.

Even a superficial observation of typical examination procedures of teachers makes it apparent, however, that the basic aims of examinations are not achieved in many instances, for the reason that the tests constructed and used by teachers fail to accomplish what is expected of them. It is indeed unfortunate that teachers, sometimes not realizing that their tests fail to accomplish the desired purposes, unduly penalize pupils for lack of success on the tests. It is necessary, therefore, that the weaknesses of classroom tests be recognized and that the proper steps be taken to bring teacher-made or classroom tests to as high a level of efficiency as possible.

## 2 ORAL EXAMINATIONS

Important as the oral quiz may be for instructional purposes, little need here be said concerning its use in the classroom for measurement purposes. When used as a teaching device in the Socratic manner, as a method of leading pupils by astute questioning to the



attainment of new understandings, oral questioning has teaching but not measuring significance. As a fact-finding technique in the interview, and in questioning the individual pupil on specific aspects of his work for obtaining diagnostic leads, oral questioning has evaluative possibilities. These, however, are not situations of the type in which oral examinations have most typically been used in the attempt to measure pupil achievement. As was pointed out earlier in this volume, Horace Mann sounded the death knell for such a use of the group oral examination more than a century ago.<sup>1</sup> Experimental evidence from many studies has indicated that the oral examination of individuals is seriously lacking in reliability and validity.<sup>2</sup>

### Limitations of the oral examination

To summarize Horace Mann's statements or implications, the oral examination: (1) is not equally fair and just to all pupils, (2) does not test extensively or efficiently, (3) permits interference and favoritism, intentional or otherwise, by the teacher, (4) is unjustifiably time-consuming, (5) leaves no permanent objective record of pupil performance, and (6) does not permit an evaluation of the difficulty of questions. While these indictments by Horace Mann accomplished very little in the sense of effecting any immediate widespread changes in examination practices, the weaknesses of the oral examination for measurement purposes have probably not since been stated more effectively.

### Advantages of the oral examination

The oral examination or quiz does have some uses, however, in the evaluation and measurement of pupil performance, even though its values admittedly are not great when it is used in the classroom situation. Certain types of performances, such as oral language, pronunciation in the foreign languages, and group performances in debates and glee club competitions, can be evaluated only in terms of the oral production. However, the evaluation of such perform-

<sup>1</sup> Otis W. Caldwell and Stuart A. Courtis, *Then and Now in Education*, 1845-1923. World Book Co., Yonkers, N. Y., 1923. p. 37.

<sup>2</sup> I. N. Thut and J. Raymond Gerberich, *Foundations of Method for Secondary Schools*. McGraw-Hill Book Co., Inc., New York, 1949. p. 163.

ances is quite different from attempts to measure pupil achievement by judging the quality of oral responses of a factual nature. Oral questioning can be used with an individual pupil in probing his reasons for having responded as he did to certain questions on written examinations or on certain mathematical or scientific problems, in an attempt to determine the causes of error. In this sense it is a diagnostic testing tool. Oral questioning can be used in determining how well an individual pupil has integrated his knowledge, can apply it to various situations, and sees its implications. Oral examinations may be used with individual pupils satisfactorily if proper advance preparations are made, if consistent procedures are followed in the question session, and if scoring and rating methods are systematically applied. However, this use of the oral examination is very time-consuming and highly subjective—qualities that make it impracticable for use with each pupil in a class for purposes of pupil comparisons.

In considering the above legitimate uses of oral questioning, it should be clearly noted that the conditions under which this method is properly used and the purposes it is appropriately expected to serve are very different from those operating when it is used with a group of pupils in the classroom to determine educational achievement. In general, the oral examination has relatively little utility in the classroom for measuring achievement, especially as a basis for determining pupil marks in a course.

### 3 ESSAY EXAMINATIONS

The traditional or essay examination continues to occupy an important place among the testing techniques used by the classroom teacher, although during the past few decades it has lost the dominant position it occupied at the turn of the century. Skepticism concerning the traditional examination arose more than a decade before 1900.<sup>3</sup> Edgeworth published in England during 1890 what was perhaps the first critical study of the essay test.<sup>4</sup> It remained, however, for Starch and Elliott to bring the issue sharply to the

<sup>3</sup> I. L. Kandel, *Examinations and Their Substitutes in the United States*. Carnegie Foundation for the Advancement of Teaching, Bulletin No. 28. The Foundation, New York, 1936. p. 27-35.

<sup>4</sup> F. Y. Edgeworth, "The Element of Chance in Competitive Examinations." *Journal of the Royal Society*, 53:460-75, 644ff.; 1890.



front in America in 1912 by a report of marks assigned to an English examination paper by various teachers<sup>5</sup> and to follow it shortly by similar reports on two other subject fields. Although it is probable that educators for various reasons somewhat misinterpreted the findings of these and many subsequent studies of the traditional examination, the fact remains that the studies very effectively called attention to a major weakness of this testing technique.

### Limitations of the essay examination

Two major limitations and several related minor limitations characterize the essay examination. The two major limitations of the essay examination, (1) limited sampling, and (2) subjectivity of scoring, are discussed in some detail in the following paragraphs, and the minor limitations are discussed briefly.

*Limited sampling.* The first major limitation of the essay examination is its limited sampling of the content of the course. A test that consists of five or ten questions cannot hope to sample widely over any sizable area of content or activities, but can measure only a few of the important areas in which pupil abilities should be tested.

Figure 8 shows in graphic form an hypothetical testing situation that points out sharply the undesirable results of limited sampling. Each one of pupils A, B, C, D, and E knows exactly half of the material over which the test is to be given. However, the particular facts mastered by each pupil are not the same throughout. For example, Pupil A, who was perhaps regular in his attendance during the first half of the course, has a mastery of the earlier units of the course. This is indicated by the shaded portion of the column. The second pupil, through irregular attendance, spasmodic preparations, or other unknown causes, mastered a few of the facts, missed another section, and then perhaps learned a few more. Pupil C was just as irregular in his attendance, but for some reason learned exactly those items missed by Pupil B. Pupils D and E show other variations of the situation. It might be carried on almost indefinitely, but these five cases are adequate to illustrate the entire range of variation due to sampling.

Now if a typical essay examination consisting of four ques-

<sup>5</sup> Daniel Starch and Edward C. Elliott, "Reliability of Grading High School Work in English." *School Review*, 20:442-57; September 1912.

tions from the various areas of the course is given, it will be noted from this diagram that distinctly different types of responses are secured from these pupils. Pupil A, knowing the facts in the first part of the work, responds to the first two questions and makes a score of 50 per cent. The second pupil, B, by sheer chance or unfortunate guidance in the selection of the facts he learned, misses each of the four questions, and receives a zero score. Pupil C, through good fortune (or judgment), happens to have mastered the items in the exact areas sampled by the test, and thereby makes a perfect score on the test. Pupils D and E, illustrating other variations due to chance, score 75 per cent and 25 per cent on the examination. Thus there is a variation of from 0 to 100 per cent on the examination taken by

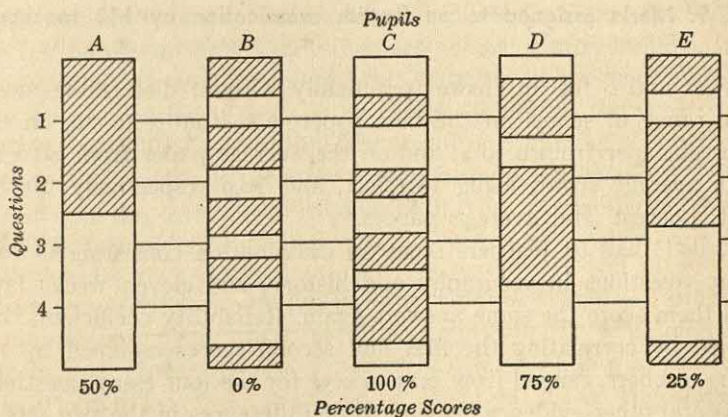


Fig. 8. Effect of limited sampling on test scores

five pupils each of whom actually has a mastery of exactly 50 per cent of the facts. This type of error in measurement of achievement, which unfortunately is not uncommon, is due to the factor of inadequate sampling. The effect of increasing the sampling from four items to many items is demonstrated by a further development and discussion of this diagram on page 164 of Chapter 7.

*Subjectivity of scoring.* A second outstanding characteristic of the essay examination is subjectivity of scoring. Starch and Elliott, who had 142 teachers score identical copies of an English examination paper, found that the scores based on 100 per cent for perfection ranged from a low of 50 to a high of 98.<sup>6</sup> In another study, they

<sup>6</sup> *Ibid.*



found that 115 teachers rated a geometry paper from a low of 28 to a high of 92.<sup>7</sup> Ruch had 91 teachers of geography score the essay examination papers judged to be the best, average, and poorest papers from a class on the basis of 20 for an entirely satisfactory

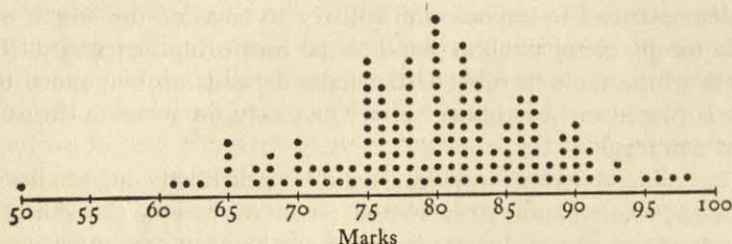


Fig. 9. Marks assigned to an English examination by 142 teachers.<sup>8</sup>

answer and 0 for an answer practically without discernible merit. The range of scores on the best paper was from 3 to 20, on the poorest paper from 0 to 2, and on the average paper from 2 to 20, with average scores being 16.1, 0.1, and 10.9 respectively for the best, poorest, and average papers.<sup>9</sup>

Eells<sup>10</sup> had 61 teachers score an examination consisting of four essay questions in geography and history, and eleven weeks later had them score the same answers again. Reliability coefficients, obtained by correlating the first and second scores assigned by the same teachers, ranged from 0.25 to 0.51 for the four essay questions. This and other evidence showing wide differences in the two sets of scores assigned by the same persons, led him to conclude that the same individuals vary from time to time in their judgments about as widely as different individuals vary.

Stalnaker, on the basis of an extensive experiment in the evaluation of English papers, concluded that "the typical essay test as typically handled... is not reliably graded and, therefore, cannot stand alone as a good measuring instrument."<sup>11</sup> From these and

<sup>7</sup> Daniel Starch and Edward C. Elliott, "Reliability of Grading High School Work in Mathematics." *School Review*, 21:254-59; April 1913.

<sup>8</sup> Starch and Elliott, "Reliability of Grading High School Work in English."

<sup>9</sup> G. M. Ruch, *The Objective or New-Type Examination*. Scott, Foresman and Co., Chicago, 1929. p. 78-81.

<sup>10</sup> Walter C. Eells, "Reliability of Repeated Grading of Essay Type Questions." *Journal of Educational Psychology*, 21:48-52; January 1930.

<sup>11</sup> John M. Stalnaker, "The Essay Type of Examination." *Educational Measurement*. American Council on Education, Washington, D. C., 1951. p. 499.

other investigations it becomes apparent that the scoring of the essay examination is a highly subjective process and that the resulting scores are correspondingly inaccurate.

The effect of a lack of objectivity in the unit of measurement may be demonstrated to anyone who will try to measure the length of a table top by using a rubber band as the measuring instrument. The length of the table in rubber-band units depends on how much tension is placed on the rubber band. Obviously no accurate measurement can result.

The subjectivity of scoring shown by practically all studies of the essay examination is more the result of varying standards of expectancy among the teachers concerned than of any other cause. Such standards of expectancy vary from day to day, teacher to teacher, grade to grade, and school to school. Unfortunately, from the point of view of improving the accuracy of scoring the essay test, this limitation appears to be largely innate in the type of examination itself. The establishment of uniform standards of achievement in the teacher is probably a human impossibility. The remedy lies not in the attempt to produce it but in giving the teacher a tangible unit of measurement.

TABLE 7. Shifting standards of expectancy

Quality of Products	Grade				
	4	5	6	7	8
82	A	A	A	B	C
68	A	A	B	C	D
50	A	B	C	D	F
32	B	C	D	F	F
18	C	D	F	F	F

Shifting standards of expectancy may be illustrated by the data of Table 7. Here are shown the shifts in standards that enter into teachers' estimates of school products. If it is assumed that a given school product, such as a handwriting or drawing specimen, has a rating-scale value of 50, it appears from the table that the specimen might receive a superior mark at the fourth-grade level. It would represent distinctly superior work for that grade. To the eighth-grade teacher, however, the specimen would appear to be very inferior as an eighth-grade product and a very poor mark might be assigned to it.



*Minor limitations.* Another factor which affects the teacher in his marking of examination papers, but which did not enter into the studies reported above, is that he typically knows the pupils whose papers he is marking and also ordinarily knows whose paper he is scoring at a given time. He is certain to be influenced by that knowledge. He is probably prone to give pupils who have previously done good work or at least made favorable impressions on him the advantage of the "halo effect." This term describes the tendency to give high marks to such pupils in some instances where they are not deserved, by explaining to himself, perhaps unconsciously, that he *knows* they know the correct answers even though their responses to the questions are not highly satisfactory to him. Similarly, the pupils he has catalogued as of low ability are sometimes penalized by his tendency to consider their good answers merely "shots in the dark" or as implying ideas that the pupils did not actually understand.

Still another type of factor affecting the objectivity of scoring of an essay test is found in the influence upon the reader of handwriting and general neatness of the paper; spelling, punctuation, and grammar; organization of the paper; and even its length. It is certainly true that a neatly typed paper of mediocre content receives the benefit of the doubt from most readers. It is also true that such characteristics as good handwriting, English usage, and organization predispose the reader toward high marks, and it is self-evident that the slow writer is penalized if a premium is attached to length of responses apart from their quality in examinations where the time is rigidly restricted. Some teachers penalize a pupil for deficiencies of these types, but other teachers do not. Many teachers are also unconsciously affected by the quality of a paper read just previously to the one being marked. Moreover, the same teacher may penalize a pupil for such deficiencies one day and not do so on another occasion, depending on his mood at the moment, and may penalize some pupils and not others.

Other influences which in considerable degree enter into the marking of tests are evidences of pupil effort, improvement, attitude toward the teacher and the course, conformance, and a multitude of other indications of what the teacher might consider desirable behavior on the part of the pupil. Some teachers believe in assigning relatively higher marks to pupils who try but do poorly than to pupils who appear not to try but do well. Others assign good marks

to pupils who conform to sometimes inconsequential and irrelevant demands and penalize pupils who do not conform.

The two types of scoring errors accounting for the subjectivity of the essay test are known as constant errors and variable errors. Constant errors are those that result from a tendency to mark high or to mark low, i.e., to be an "easy" marker or a "hard" marker. Variable errors result from the tendency of all persons to vary in their judgments from time to time, according to their states of mind, the states of their digestions, and many other factors.

Pupils who do not know the answers to essay questions are prone to respond in terms calculated to cover up their lack of information if not actually to mislead the teacher. Such responses, which tend to vary in plausibility directly in relation to the intelligence of the bluffer, may take the form of discussion concerning content closely related to that covered by the question, of very incomplete answers which by repetitious statements and copious illustration may give a sense of completeness, and various other devices. Whether bluffing is or is not desirable is not the issue. Certainly bluffing is resorted to in great or small degree by all persons on some, if not many, occasions. To the extent, however, that bluffing is actually successful on essay tests, the examination results are less accurate measures of pupil achievement.

Stalnaker, after presenting evidence that the cost of securing an accurate reading (evaluation) of essay questions is practically prohibitive, summarized his position on the problems of the essay examination as follows: "The accurate evaluation of a well-developed essay question is a long and difficult job and one which, properly done, requires intelligence, diligence, and consistency. The expense in time and money can be justified only to the extent that essay items are developed to measure reliably important objectives which cannot otherwise be measured."<sup>12</sup>

### Advantages of the essay examination

Only the major and rather commonly accepted advantages of the traditional or essay examination will be discussed here. It should be remembered here particularly that it is the total effect of the examination which is important rather than the specific aspects

<sup>12</sup> *Ibid.* p. 502.



considered singly. An advantage, then, may not be an advantage when there is balanced against it one or more dependent disadvantages.

*Ease of construction and administration.* Essay tests are commonly considered easy for teachers to prepare and to administer. Pupils feel that they know the nature of essay test questions and the traditional methods of answering them. Teachers typically give a minimum of time to the preparation of essay questions: Sometimes the questions are not even formulated until immediately before the examination is to be given. Some teachers even prepare the last part of the test while the pupils are writing on the first questions. Little or no time is taken for telling pupils how to take the essay examination. However, essay tests prepared and administered with a minimum of effort are likely to have such resulting disadvantages that the saving of time and labor may well be at the expense of testing efficiency.

*Adaptability to school subjects.* It is possible to use the essay examination for practically all subjects of the school curriculum, for the question and answer method is widely adaptable. Some types of outcomes, such as arithmetic skills, handwriting skills, reading ability, and others, cannot be tested directly by this device, but the factual backgrounds for them frequently can be so tested. As a matter of fact, the essay test procedure is often used in scoring arithmetic examples by the use of arbitrary decisions in scoring for correctness of the result or correctness of the method, in giving partial credit for answers not entirely correct, and in various other ways.

*Measurement of higher mental abilities.* Advocates of the older type of examination insist that the discussion-type questions have values not possessed by the informal objective test in that they call for comparison, for interpretation of facts, for criticism, for defense of opinion, and for other types of higher mental activity. Essay questions allow for some range of choice, which makes possible the meeting of differences in courses and readings pursued. The purpose of the written test is primarily to ascertain whether the student has accurate knowledge and a considerable amount of understanding about a wide variety of matters in terms of their interacting relationships but not basically to determine whether he knows certain highly detailed facts and whether he has met routine course re-

quirements. What is sought is a measure of accurate knowledge of fact, understanding of complex ideas, and ability to interpret and to criticize and decide. In short, the questions are devised to test the pupil's ability to make use of knowledge. This is particularly true for advanced students, for whom the testing of such types of higher abilities is more important than the testing of the broad factual knowledges that almost certainly have been acquired to a high degree.

### Advantages claimed for the essay test

Various advantages have been claimed for the traditional examination. Some of these advantages appear to depend on evidence that is not too conclusive. In many cases the decision depends as much on the philosophy of the individual teacher as on definite research findings, so that possible advantages cannot be claimed with certainty.

*Freedom of response.* The freedom of response that the essay test question allows is considered by some students of examination methods as one of its fundamental characteristics. By the nature of the question the student is required to survey his own background of related information and to select the related facts and organize them for expression in his own words. It is important, however, that the freedom of selection, organization, and expression be suited to the measurable outcomes of the course.

*Training in the use of written English.* It has been contended by various persons in the past that training in the effective use of English is a logical function of the examination and that the essay test furnishes such training. However, neither contention is defensible. Courses in English provide training in the use of English, as do, indirectly and as by-products, many other types of school experiences. The examination, which has definite uses and purposes in the measurement area to occupy its attention, should not be expected to furnish training in the use of English, although, of course, written language is required in the essay test. Furthermore, the conditions under which the essay examination is typically given—pupils writing at high speed and without time to organize their thoughts carefully—are not conducive to the best use of English. Certainly examinations in such courses as language, composition,



spelling, and perhaps reading and literature might be devised to furnish training in the written use of English, but there seems to be no justification for shortening the time given to the measurement of direct course outcomes in such subjects as the sciences, social studies, and arithmetic in order to furnish the pupils this type of training.

*Motivation of desirable methods of reviewing.* Many teachers feel that some student groups prepare for essay tests more often by reviewing broadly the important aspects of course content but that they more frequently review for the objective test by memorization of facts or exact wordings of the textbook. No one would deny the general desirability of the first rather than the second type of review. However, such opinions are usually based on observations of how a few groups of pupils say they prepare for examinations. Probably the type of examination is less important to the pupil in determining how he should review than the nature of the test. An essay examination may or may not stress detailed facts. An objective test may or may not stress detailed facts. Teachers differ markedly in the emphases they assign to factual learning and to applications of facts in the tests they give.

## Conclusions concerning the essay examination

For many years the essay-type test has been subjected to intense criticism. In spite of these attacks, however, it is still in use in numerous classrooms and doubtless performs a worthwhile function there. While it is true that when the essay test is subjected to a critical appraisal under research conditions many of the claims that have been advanced for it do not stand up any too well, it is also true that it performs certain functions in the classroom and for the pupils that certain of the other more objective forms of tests fail to accomplish. Without doubt the essay-type test is firmly fixed in educational practice. It is a type of examination with which all teachers are familiar, and with all of its faults it undoubtedly possesses sufficient merit to warrant considerable attention to its improvement.

It is now recognized that only a portion of the variability of marks assigned to an examination by *different* teachers, as in the Starch and Elliott and other studies, can be attributed to the un-

reliability of the essay examination itself. A comparable share of the variability can be charged to the lack of uniformity in the scoring procedures followed by the teachers. Whereas the different teachers used in such studies had very different educational aims and standards of excellence, the teacher who scores an entire set of papers attempts to apply the same set of standards to all papers, and has the benefit of experience with previous papers as a basis for doing so. Furthermore, the teachers in those studies used no scoring rules save those which they developed individually, but the teacher who scores a set of papers usually applies more or less tangible and consistent scoring procedures.

A final summation of the limitations and advantages of the essay examination cannot be conclusive. Certainly the limitations of the test as it has been, and perhaps even today is, most widely used greatly outweigh its advantages. However, it may be that when the essay test is used with optimum efficiency and for carefully defined purposes many of its advantages will be realized.

#### 4 IMPROVING ESSAY EXAMINATIONS

Many suggestions for improving the essay examination have been made by students of this type of test. Most of the suggestions have to do with: (1) the selection of test content and the framing of questions, and (2) the scoring of test results. The discussion below presents a few of the approaches to the improvement of the essay test by these two methods, but does not attempt to consider how the test may be improved in the specific subjects of the school curriculum.

#### Types of essay questions

Monroe and Carter classified essay-type questions with respect to the mental activity each type is designed to elicit in the pupil, and presented both descriptive statements concerning, and examples of, the twenty varieties they distinguished.<sup>13</sup> The descriptive state-

<sup>13</sup> Walter S. Monroe and R. E. Carter, *The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students*. Bureau of Educational Research Bulletin, No. 14. University of Illinois, Urbana, 1923.



ments and illustrative questions below are from Odell's adaptation and supplementation<sup>14</sup> of the questions from Monroe and Carter's list.

1. Selective recall—basis given. (Name the presidents of the United States who had been in military life before they were elected.)
2. Evaluative recall—basis given. (Name the three statesmen who have had the greatest influence on economic legislation in the United States.)
3. Comparison of two things—on a single designated basis. (Compare Eliot and Thackeray as to ability in character delineation.)
4. Comparison of two things—in general. (Contrast the life of Silas Marner in Raveloe with his life in Lantern Yard.)
5. Decision—for or against. (In which in your opinion can you do better, oral or written examinations? Why?)
6. Causes or effects. (Why has the Senate become a much more powerful body than the House of Representatives?)
7. Explanation of the use or exact meaning of some phrase or statement in a passage. (Explain the meaning of the expression "Sinai's climb" in the line: "We Sinai's climb and know it not.")
8. Summary of some unit of the text or of some article read. (Summarize in about one hundred words the advantages of the hot-air furnace.)
9. Analysis. (Mention several qualities of leadership.)
10. Statement of relationships. (Tell the relation of exercise to good health.)
11. Illustrations or examples (your own) of principles in science, construction in language, etc. (Give an original sentence in Latin illustrating the use of the infinitive in indirect discourse.)
12. Classification—usually the converse of No. 11. (To what group of plants do the mosses and liverworts belong?)
13. Application of rules or principles in new situations. (In what countries other than Brazil would you expect to find rubber plantations?)
14. Discussion. (Discuss the Monroe Doctrine.)
15. Statement of aim—author's purpose in his selection or organization of material. (What was the purpose of the author in having Athelstane return to life after he was apparently dead?)
16. Criticism—as to the adequacy, correctness, or relevancy of a printed statement, or a classmate's answer to a question on the lesson.

<sup>14</sup> C. W. Odell, *Traditional Examinations and New-Type Tests*. Century Co., New York, 1928. p. 207-10.

- (Criticize "Macbeth was wholly indifferent to the superstitions of his time.")
17. Outline. (Outline in not more than one page the chief events of the French and Indian Wars.)
  18. Reorganization of facts. (Select the incidents which characterize Portia in *The Merchant of Venice*.)
  19. Formulation of new questions—problems and questions raised. (If you were asked to state how much you could trust the viewpoint of a particular historian about whom you know little or nothing, what questions would you want to have answered concerning him?)
  20. New methods of procedure. (How might the plot of *Julius Caesar* be changed to make it a comedy rather than a tragedy?)

Questions of the essay type are commonly classified into three types: (1) simple-recall, (2) short-answer, and (3) discussion. The simple-recall questions, demanding a short response that can be accurately scored, require a name, a number, a date, a place, or an event in answer to *who*, *how many*, *when*, *where*, and *what* questions. The short-answer questions, demanding statement, phrase, or sentence responses that can be rated quite objectively, require answers to such key words as *define*, *identify*, *list*, *find*, and *state*. The discussion questions, requiring responses of such complexity that objectivity of scoring is difficult, request answers to such words as *discuss*, *explain*, *describe*, *compare*, and *outline*. As most teachers are well aware, some essay questions are sufficiently definite that responses can be evaluated objectively, but others are so general that responses can be rated with reasonable accuracy only by the use of definite scoring rules or some similar method.

### Increasing the objectivity of scoring the essay test

Approximately forty years ago, Kelly conducted an investigation into the causes of variation in teachers' marks on examination papers.<sup>15</sup> He found that the use of a rather definite set of rules resulted in greatly reduced variations in scores when the papers were rescored. More recently, Stalnaker obtained reliability coefficients ranging from .84 to .99 for the scores assigned to essay exami-

<sup>15</sup> Fred J. Kelly, *Teachers' Marks*. Contributions to Education, No. 66. Teachers College, Columbia University, New York, 1914.



nations in a variety of high-school subjects by experienced teachers when scoring rules were used.<sup>16</sup> These reliability coefficients show a highly satisfactory degree of scoring accuracy, especially when it is considered that only the lowest coefficient was under .90. Other studies of the results obtained when the essay test was scored under closely controlled conditions substantiate the conclusion that the traditional examination can be scored reliably if proper precautions are taken.

Sims proposed a rating method of scoring essay examinations.<sup>17</sup> He suggested that the readers work out for themselves acceptable answers to the questions and then use the following procedures:

- a. Quickly read through the papers and on the basis of your opinion of their worth sort them into five groups as follows: (a) very superior papers, (b) superior papers, (c) average papers, (d) inferior papers, (e) very inferior papers. (Remember that in a normal group you would expect to have approximately 10 per cent of *very superior* and 10 per cent of *very inferior* papers, 20 per cent of *superior* and 20 per cent of *inferior* papers, and 40 per cent of *average* papers. Do not, however, try to conform rigidly to this rule. Your group may not be a normal one.)
- b. Re-read the papers in each group and shift any that you feel have been misplaced.
- c. Make no attempt to give numerical grades or to evaluate each question. Place each paper on the basis of your general impression of the total.
- d. Assign letter grades to each group; beginning with A for the very superior group, B for the superior group, etc.

Stalnaker reported evidence that the use of optional questions appears to reduce the reliability of marking the essay examination and recommended that optional questions be avoided.<sup>18</sup>

Wrightstone recommended that essay tests be designed to measure only one objective of instruction at a time, such as interpretation of facts, that all scorers agree on a definition of the objectives and on

<sup>16</sup> John M. Stalnaker, "Essay Examinations Reliably Read." *School and Society*, 46:671-72; November 20, 1937.

<sup>17</sup> Verner M. Sims, "The Objectivity, Reliability, and Validity of an Essay Examination Graded by Rating." *Journal of Educational Research*, 24:216-23; October 1931.

<sup>18</sup> John M. Stalnaker, "The Essay Type of Examination." *Educational Measurement*. American Council on Education, Washington, D. C., 1951. p. 506.

certain standards of values, that an ideal answer be formulated and each part assigned a certain number of points, and that an eleven-point scale from 0 to 10 be used for each test unit.<sup>19</sup>

The following suggestions, by largely eliminating the personal judgment or bias of the scorer, have been found valuable for use in scoring essay-type responses:

1. Examinations should be scored by the one who makes out the questions. He should know exactly what responses are desired, and should write out his answers to the questions in advance.
2. Each pupil taking the test should write his name on the back of the test paper and the scorer should disregard the name until the test is scored. This eliminates the subjective factor of being influenced or biased in judgment because of former contacts with the pupil, insofar as the teacher does not become aware of the writer's identity through his handwriting or his manner of expression.
3. The scorer should not mark off for misspelled words or poor sentence structure, paragraphing, or handwriting. Similarly, he should not increase the score for excellence in these things. However, such factors may be indicated or checked on the examination. The reason for this lies in the fact that the function of the examination is to measure the pupil's abilities in a course and not his ability to write or to spell. If it is desirable to test his ability to write, spell, or use correct written English, suitable tests can be obtained for these purposes.
4. Each separate item should be scored in all of the papers consecutively. This is preferable to the correction of each entire test as a unit, for it permits the scorer to concentrate on the answer to a single test question and to judge better the merits of the several pupil responses to the same question.
5. Each question should be rated on a scale of ten, twenty, or a given number of scoring points. The total score should be obtained for each pupil by adding the scores on the different questions only after all of the scoring has been done.

Whatever rules are followed, they will necessarily be arbitrary and not always wholly defensible. The significant point in the use of rules is that they provide for reasonable uniformity in handling the papers of all the pupils and also furnish a guide for the control of irrelevant factors that may affect the objectivity of the scoring.

<sup>19</sup> J. Wayne Wrightstone, "Are Essay Examinations Obsolete?" *Social Education*, 1:410-15; September 1937.



## Suggestions for improving the essay examination

Four conditions appear to be necessary in bringing about improvement in the teacher-made examination of the essay type. These conditions are:

- (1) *The exact purpose of the examination must be understood by both the teacher and the pupil.* The emphasis of the essay examination should be definitely on thought, reasoning, and other types of mental activity as applied to the materials of the course. The main concern is with topics that involve interest-centers or relationships and problematical issues. Answers to questions involving judgments, synthesis, and generalizations are admittedly difficult to evaluate, but they reveal aspects of pupil mastery and mind quality probably not obtained from other types of responses.
- (2) *The content of the examination should be governed by its purpose.* In general, a test should parallel the objectives and pupil outcomes of the course. This means that there should be a proper balance of test content not only with respect to the subject matter but also with respect to the types of abilities to use and apply informations that are desired pupil outcomes.
- (3) *The preparation and selection of suitable essay-type questions should consume at least as much time as is required to score the answers.* If this is done, the value and the accuracy of the scores obtained are almost certain to be increased.
- (4) *Definite rules should be formulated that will as far as possible control irrelevant factors in scoring the papers.* The careful use of scoring rules will bring about a definite decrease in the inaccuracy of the pupil scores.

The accompanying tentative score card for rating essay-type questions is suggested as a possible means of improving this type of teacher-made examination by calling attention to the desirable qualities in test questions. Unless a question rates "Yes" on at least seven of the ten items, it is certainly of doubtful value and should probably be rewritten and given a new emphasis or be completely eliminated from the examination.

## Tentative Score Card for Rating Essay-Type Examination Questions

	Yes	Slightly	No
1. Is the question concerned with important phases of the subject? .....			
2. If the question emphasizes minor details, are they useful in linking up other facts, ideas, theories, involved in the subject? .....			
3. Does the question give emphasis to evaluation and to relational thinking? .....			
4. Is the question apparently of a suitable degree of difficulty in relation to the other questions in the test? .....			
5. Is the question stated in such a way as to stimulate thought, to challenge the interest of the pupils? .....			
6. Does the question motivate the pupil to integrate his ideas around certain interest-centers? .....			
7. Is the question stated in such form as to cause the pupil to sample widely into his background of fact? .....			
8. Does the question call for any originality of thought organization and expression? .....			
9. Does the question call for the pupil to integrate facts gained from different sources? .....			
10. Is the question limited sufficiently that the pupil has some chance of writing what he really knows about it in a reasonable time? .....			

## Topics for Discussion

1. Indicate why there is need for improvement in classroom testing.
2. What are some of the major weaknesses of the oral examination for testing purposes?
3. What uses should the oral quiz be expected to serve in the school?
4. Discuss fully the manner in which limited sampling reduces the reliability of the essay examination.
5. List and discuss several factors that contribute to subjectivity of scoring the typical essay examination.
6. Comment upon some of the minor weaknesses of the essay test.
7. List and evaluate the advantages that have been attributed to the traditional examination.



8. What are your conclusions concerning the proper place of the essay test in classroom measurement?
9. Identify some of the types of essay questions and indicate key words by which they are introduced.
10. Suggest at least five specific devices or procedures for increasing the objectivity of scoring essay-type tests.
11. Outline testing procedures by which the essay-type test may be made more effective as a classroom testing technique.

### Selected References

- ASHBURN, ROBERT. "An Experiment in the Essay-Type Question." *Journal of Experimental Education*, 7:1-3; September 1938.
- BARNES, ELINOR J., AND PRESSEY, S. L. "The Reliability and Validity of Oral Examinations." *School and Society*, 30:719-22; November 23, 1929.
- BENDER, WILLIAM, JR., AND DAVIS, ROBERT A. "What High School Students Think about Teacher-Made Examinations." *Journal of Educational Research*, 43:58-65; September 1949.
- BRODY, WILLIAM, AND POWELL, NORMAN J. "A New Approach to Oral Testing." *Educational and Psychological Measurement*, 7:289-98; Summer 1947.
- CALKINS, MARY W. "Philosophers in Council." *School and Society*, 17:316-20; March 24, 1923.
- CASON, HULSEY. "An Intelligence-Question Method of Teaching and Testing." *Pedagogical Seminary and Journal of Genetic Psychology*, 54:359-90; June 1939.
- ENGELHART, MAX D. "Examinations." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 407-14.
- GERBERICH, J. RAYMOND. "Tests and Measurement." *Review of Educational Research*, 15:408-22; December 1945.
- GREENE, HARRY A., AND CRAWFORD, JOHN R. *Work-Book in Educational Measurements and Evaluation*. New York: Longmans, Green and Co., 1945. Unit 1.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. p. 41-43, 57-65.
- KANDEL, I. L. *Examinations and Their Substitutes in the United States*. Carnegie Foundation for the Advancement of Teaching, Bulletin No. 28. New York: The Foundation, 1936.
- LAWSON, D. E. "Historical Survey of Changes in Aims and Outcomes of School Examinations." *Educational Administration and Supervision*, 26:667-78; December 1940.

- MEYER, GEORGE. "The Choice of Questions on Essay Examinations." *Journal of Educational Psychology*, 30:161-71; March 1939.
- MONROE, WALTER S., AND CARTER, R. E. *The Use of Different Types of Thought Questions in Secondary Schools and Their Relative Difficulty for Students*. Bureau of Educational Research Bulletin, No. 14. Urbana: University of Illinois, 1923.
- ODELL, CHARLES W. *How To Improve Classroom Testing*. Dubuque, Iowa: Wm. C. Brown Co., 1953. Chapter 5.
- ODELL, CHARLES W. *The Use of Scales for Rating Pupils' Answers to Thought Questions*. Bureau of Educational Research Bulletin, No. 46. Urbana: University of Illinois, 1929.
- REMMERS, H. H., AND GAGE, N. L. *Educational Measurement and Evaluation*. New York: Harper and Brothers, 1943. Chapter 12.
- ROSS, C. C. *Measurement in Today's Schools*. Second edition. New York: Prentice-Hall, Inc., 1947. Chapter 6.
- SIMS, VERNER M. "The Essay Examination Is a Projective Technique." *Educational and Psychological Measurement*, 8:15-31; Spring 1948.
- SIMS, VERNER M. "Essay Examination Questions Classified on the Basis of Objectivity." *School and Society*, 35:100-2; January 16, 1932.
- STALNAKER, JOHN M. "The Essay Type of Examination." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 13.
- TRAXLER, ARTHUR E. "A Note on the Accuracy of Teachers' Scoring of Semi-Objective Tests." *Journal of Educational Research*, 37:212-13; November 1943.
- TRIMBLE, OTIS C. "The Oral Examination: Its Validity and Reliability." *School and Society*, 39:550-52; April 28, 1934.
- VALLANCE, THEODORE R. "A Comparison of Essay and Objective Examinations as Learning Experiences." *Journal of Educational Research*, 41:279-88; December 1947.
- WEIDEMANN, CHARLES C. "Review of Essay Test Studies." *Journal of Higher Education*, 12:41-44; January 1941.
- WEIDEMANN, CHARLES C. "Further Studies of the Essay Test." *Journal of Higher Education*, 12:437-39; November 1941.



## ***Constructing and Using Informal Objective Tests***

THIS CHAPTER deals with the following points concerning the construction and classroom use of informal objective tests:

- A. Similarities of the informal objective examination and the standardized test.
- B. Major advantages and limitations of the teacher-made objective examination.
- C. Types of instructional outcomes.
- D. Selecting the content and preparing an informal objective test.
- E. Administering and scoring the informal objective test.
- F. Uses and limitations of basic objective item forms.
- G. Illustrations of objective-test item types.
- H. General suggestions for constructing objective test items.
- I. Suggestions for constructing basic types of recall and recognition test items.

Developments contributing to the improvement of measurements in education have largely followed two main lines: (1) the construction and improvement of standardized tests, and (2) the improvement of teacher-made tests. It is with the second of these that this chapter is concerned. In many respects these two types of measurement are not fundamentally different. Both utilize samplings of material to stimulate pupil reactions. In both the performance is expressed in terms of a score. Both make use of exercises that are

characterized by being objective. Yet, in spite of these similarities, the two types of tests do not seriously overlap in function.

## 1 CHARACTERISTICS OF CLASSROOM TESTING

### Importance of classroom testing

Even though standardized instruments for measuring achievement of school children have come into wide use, the examination constructed by the teacher still remains the most frequently used means of measuring the achievement of pupils. Although properly constructed standardized educational tests may be superior in certain respects to teacher-made examinations, they should never displace the teacher-made test as a means of measuring the results of teaching as indicated by pupil attainment. The teacher frequently has need for a measuring instrument adapted to a particular course of study or to the instructional emphasis that has been given to the subject in the teaching of a particular class. The informal objective examination constructed by the teacher to fit the instruction the class has been receiving is the obvious answer.

### Standardized vs. non-standardized objective tests

Standardized educational tests are structurally not fundamentally different from informal objective examinations in their basic elements. In fact, standardized educational tests are essentially little more than improved and refined objective examinations.

In contrast with their similarities from a structural point of view, the functions of the standardized test and the informal objective examination over the same material are quite distinct. The standardized test, because it is intended for use in many different school systems and in connection with many different types of courses of study, must be general as to content. The maker of a standardized test cannot be sure that its content will actually parallel the instructional emphasis given the subject in the course offered by any individual teacher. Accordingly, the standard test is useful mainly for general comparisons of school with school, class with class, or city with city. It is not designed for use in evaluating the accomplishment of pupils in a class under a particular instructor with a specialized instructional emphasis. By the same reasoning, *the standardized test should prob-*



*ably not be used as the basis for the assignment of course marks in any subject.*

The informal objective examination, constructed in accordance with well-recognized principles and incorporating extensive samplings of the content actually taught by the teacher and the activities of his pupils, is, on the other hand, a suitable basis for the assignment of such course marks. It is quite probable that even though two objective tests, one standardized and one informal, could be made equal in objectivity, length (in terms of number of items as well as testing time), and reliability of measurement, their functional values in the classroom would still be quite unlike, because of unavoidable differences in their content alone. Thus, in general, standardized and informal objective tests must be considered as having quite distinct and separate functions, and the terms are not to be used interchangeably.

Tyler was the leader some twenty years ago in a movement to broaden the base for informal objective testing. He pointed out that test content had been validated primarily in terms of the informational content of the courses tested, and recommended a procedure that validated test content in terms of course objectives. Tyler's recommendations for procedures to be followed in achievement test construction are reproduced below without discussion at this point.<sup>1</sup> Recent enlightened attacks upon construction of both informal objective tests and standardized tests have doubtless been influenced significantly by this point of view.

1. Formulation of course objectives.
2. Definition of each objective in terms of student behavior.
3. Collection of situations in which students will reveal presence or absence of each objective.
4. Presentation of situations to students.
5. Evaluation of student reactions in light of each objective.
6. Determination of objectivity of evaluation.
7. Improvement of objectivity, when necessary.
8. Determination of reliability.
9. Improvement of reliability, when necessary.
10. Development of more practical methods of measurement, when necessary.

<sup>1</sup> Ralph W. Tyler, "A Generalized Technique for Constructing Achievement Tests," *Educational Research Bulletin*, 10:199-208; April 15, 1931.

## 2 ADVANTAGES AND LIMITATIONS OF INFORMAL OBJECTIVE TESTS

The foregoing discussion has pointed out that the standardized and the informal objective test are closely similar in the form in which the test items are stated. In fact, both types of examinations make use of the same general principles in the formulation of their content. Both the standardized and the non-standardized tests may include enough items to afford consistent measurement. On the other hand, there are a few very distinct differences between the essay examination and the teacher-made objective examination. In general, the advantages of the informal objective test are in the areas in which the essay test has definite limitations and perhaps to a less extent the weaknesses of the informal objective test are in the areas where the essay test is relatively satisfactory. Therefore, the treatment below is related to that of Section 3 of Chapter 6 and will in some instances depend upon the previous discussion.

Because of the similarities between the teacher-made objective test and the standardized test noted above, the treatment of test construction here applies almost equally well to both forms of objective test. Their major differences lie in the purposes for which they are constructed and in the uses to which they are properly put, whereas the discussion here is based more on the form of the types of tests being contrasted.

### Advantages of informal objective tests

Of the several merits of the informal objective test, the two most important are answers of the early objective testers to the two major criticisms of the essay examination discussed above—limited sampling and subjectivity of scoring.

*Extensive sampling.* Although all tests measure only samples of pupil performance, the objective test by its nature samples so widely that the results obtained from its use closely approximate those that would be obtained if pupil performance in the subject in question could be measured completely. A test made up of a hundred or so short, well-selected questions or items will adequately sample pupil achievement for many purposes.



The results from administering a test consisting of many items of narrow range are shown in Figure 10. This illustration is based on the same hypothetical situation as that of Figure 8 on page 143, and the same basic conditions apply. The shaded portions of the five rectangular areas represent the portions of the total course content mastered by Pupils A, B, C, D, and E. In each instance the shaded portion is exactly half of the area of the rectangle. The close-ruled horizontal lines represent the 20 short-answer questions, which are so distributed as to cover the content of the entire course.

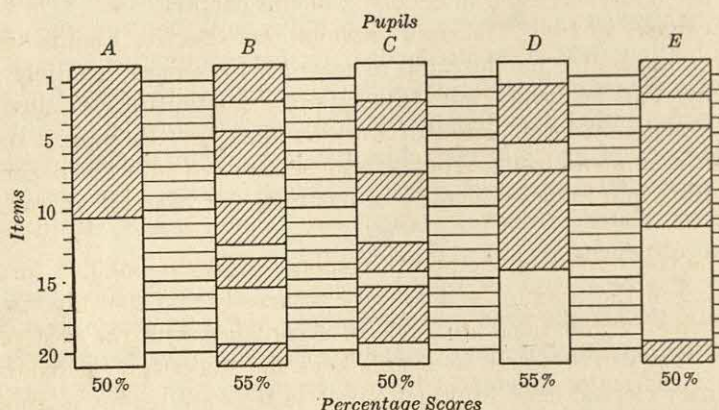


Fig. 10. Effect of extensive sampling on test scores

It is apparent here, in contrast with the results shown in Figure 8 when only four questions were used, that the five pupils receive scores that are very similar. Since exactly half of the twenty lines are opposite the shaded areas for Pupils A, C, and E, they receive scores of 50 per cent. As eleven of the twenty marks are opposite the shaded areas for Pupils B and D, they receive scores of 55 per cent. These results show that the objective test samples widely, and that scores resulting from its use are not likely to be much affected by differences in the knowledge of different pupils. Enough different questions are asked to make sure that the mark made by each pupil will place him quite accurately in relation to his classmates in terms of his knowledge of course content. This is in direct contrast to the findings based on the illustration of Figure 8, which were that wide differences occurred in the scores assigned to the five pupils.

*Objectivity of scoring.* In an objective test the items are so stated that the answers are brief, and usually only one correct answer is possible. A highly objective test may be scored repeatedly by one person or it may be scored by a large number of different persons with practically no disagreement in the scores assigned. Thus in the objective examination the responses can be evaluated on an impersonal basis, entirely independent of the personal judgment of the examiner. This is true, of course, only when the items are constructed in accordance with certain recognized principles. These principles are listed and discussed in Section 5 of this chapter.

*Economy of time.* The form in which the objective item is stated makes it possible for the pupil to record his response definitely and briefly. This in turn permits many specific reactions to be called for in a relatively brief period of working time. In this way a much wider area of the course content can be sampled in a given period, resulting in a higher reliability of measurement per unit of working time.

The conciseness of the pupil's response makes it possible for the scoring of the tests to be done very accurately and speedily. If the objective examinations are made in accordance with the best practices, they can be scored by simple keys and stencils in the hands of ordinary clerical help. Informal objective tests are readily adaptable to machine scoring.

*Elimination of bluffing.* Fluency of expression and mastery of the language have always been recognized as factors in examinations of the discussion type. Because of the nature of the items, the amount of writing done by pupils in responding to objective tests is reduced to a minimum, however. This practically eliminates bluffing and the advantage that rapid and fluent writers have over those not so gifted. The fact that one pupil can write more material than another in the same length of time should not result in his receiving higher marks in his school subjects.

### Possible disadvantages of informal objective tests

A number of rather important criticisms of objective examinations have been brought forward by teachers and critics. The following list, while not complete, probably contains the more significant of these objections.



*Neglect of training in expression of thought.* Teachers sometimes feel that the informal objective test inadequately allows opportunity for the pupil to organize and express his thoughts. One approach to this criticism is through an analysis of how well the essay test fulfills these purposes. The amount of such training that is derived from writing an essay examination is negligible at best. The time stress under which the pupil typically writes his examinations gives him very little opportunity carefully to think his way through what he actually knows about the subject. He has almost no time to consider sentence structure, paragraph organization, or vocabulary choice. The net result is that he forms bad rather than good habits of thought, expression, and work.

Some objective methods are available for testing the ability of the pupil to organize his thoughts, but no claim should be made that the objective test provides opportunity for the verbal expression of organized thought. The written examination should be expected to serve no such purpose. The opportunity for training the pupils in self-expression can and should be provided adequately elsewhere in the school program.

*Overemphasis on factual knowledge.* This objection to the objective examination overlooks the fact that almost uniformly essay questions test memory for the factual aspects of the subject. The thought question as a type is not at all inherent in the essay examination. Furthermore, there is nothing in the objective form which makes impossible the construction of items that stimulate critical and constructive thought. Many teacher-made tests do not contain such thought-provoking items, but that does not mean that they cannot be made to do so when teachers become masters of objective techniques and learn to think deeply enough into the validation of their tests. The informal objective examination can be used, as is brought out later in this chapter and in subsequent chapters of this volume, in the measurement of various instructional outcomes of significance far beyond the acquisition of facts and of basic skills. It is probable that one source of this criticism lies not so much in objective methods of measurement in general as it does in the kinds of objective material typically prepared by the individual teacher.

*Encouragement of guessing.* Some teachers and critics believe that there is a tendency for the objective test to encourage guessing to an undue extent. The objective examination form admittedly permits, but does not necessarily encourage, guessing. In fact, it may

tend to discourage guessing through its emphasis upon exact knowledges and correct applications and interpretations of factual data, and in its use of correction for guessing formulae in scoring test results. Furthermore, it is probable that few guesses on objective tests are based on pure chance. Rather they are based on slight balances of evidence on one side or the other of an issue on which the pupil is uncertain. Many life activities, as a matter fact, are based on chances considerably less than certain of a given outcome. Therefore, it seems that guessing in the sense of weighing available evidence and making the best decision possible is neither injurious to the pupil nor a bad influence upon examination results.

*Difficulty of preparation.* The criticism that informal objective tests are difficult to prepare is frequently made. The typical essay test is easy to prepare but difficult to score. The informal objective test may be difficult to prepare but it is certainly easy to score. When the advantages accruing to the use of objective tests are balanced against the difficulty of preparing them, the conclusion seems favorable rather than otherwise to the objective test.

*Considerable cost.* Experience in the use of objective examinations indicates that they are most valuable when available for classroom use in printed or mimeographed form. Unquestionably the paper cost is an item of expense which in some school systems may be serious. However, some kind of paper must be used for the examination. Mimeograph paper is approximately as cheap as any. If the teacher is willing to do his own mimeographing or hektographing, the extra expense should not be very great. As a matter of actual fact, the cost of preparing objective examinations probably represents one of the very minor items of expense in the average school system when it is considered in terms of the real educational importance of such equipment.

### 3 CONSTRUCTION AND USE OF INFORMAL OBJECTIVE TESTS

The problems of constructing and using informal objective tests are discussed in this and the following section of this chapter. Treated here are the general issues that should receive consideration from the time a test is in the planning stage to the time when its results have been used finally in the validation of its individual items. The following section deals with the various major objective item forms somewhat in detail and presents samples of the various item types.



Section 5 gives general and specific suggestions for drafting items of the five basic types.

## Types of instructional outcomes

A classification of instructional outcomes by major types is useful in obtaining an optimum balance of test content and in making sure that certain types of outcomes are not overstressed at the expense of other and perhaps even more fundamentally important outcomes. The types of outcomes<sup>2</sup> which may be distinguished are: (1) skills, (2) knowledges, (3) concepts, (4) understandings, (5) applications, (6) tastes and preferences, and (7) adjustment.

Tastes and preferences are best represented by attitudes, interests, and appreciations. The individual's feelings and emotions are much more involved in formulating his tastes and preferences than are the more largely intellectual processes operative in determining his knowledges, concepts, and understandings. Whether adjustment should be considered as a type of instructional outcome or the resultant of all of the other types of outcomes is not certain. In any event, since attitudes, interests, and adjustment are ordinarily considered to fall in the area of personality measurement, they are discussed in Chapter 11. Appreciations, not directly measurable by the usual paper-and-pencil achievement tests but rather subject to appraisal by the use of more complex evaluative tools and techniques, are dealt with in Chapter 9.

The five remaining types of outcomes—skills, knowledges, concepts, understandings, and applications—are briefly characterized below. It seems certain that care in attaining an optimum balance among these types of outcomes in test construction will result in more valid informal objective tests than might otherwise be attained. It should be borne in mind, however, that subject areas and even specific subjects differ widely in their objectives and hence in the behavioral outcomes to be expected of pupils. For example, such tool subjects as mathematics, the expressive language arts, and the practical arts appropriately emphasize skill outcomes more than do such content subjects as the social studies and the sciences. Knowl-

<sup>2</sup> Adapted from Asahel D. Woodruff, *The Psychology of Teaching*, Third edition. Longmans, Green and Co., New York, 1951. Chapter 16. See also I. N. Thut and J. Raymond Gerberich, *Foundations of Method for Secondary Schools*. McGraw-Hill Book Co., Inc., New York, 1949. p. 107-12.

edges, concepts, and understandings appropriately receive somewhat more emphasis in the content subjects than in the skill areas. The consequence of such differences is that an optimum balance among these types of outcomes in test construction is attainable only by careful consideration of the appropriate objectives and hence of the appropriate instructional outcomes separately for each course and in the usual case individually by each teacher.

*Skills.* Physical or motor activity is definitely involved in many of the types of behavior considered under this heading, although this does not preclude the operation of mental processes in skill behavior. Reading skills, work-study skills, language skills, computational skills, shop and laboratory skills, typing skills, and athletic skills are representative of the variety and of the scope covered by this type of outcome.

*Knowledges.* Attainment of knowledge involves the establishment of such mental associations as those between an object and its name, a date and an event, a term and the color or characteristic it represents, and a symbol and its meaning. Outcomes of this type are represented by knowledges concerning facts, principles, and laws, knowledges concerning processes and procedures, and knowledges concerning sources of information.

*Concepts.* Concepts presuppose that meaning has been attached to what has been learned, whereas purely knowledge outcomes may be, but are not necessarily, organized at the conceptual level. Abilities to give the meanings of words, to discriminate types or qualities of color, and to use abstract words in thinking, speaking, and writing demonstrate the attainment of concepts. The emphasis in modern schools on the development of meanings represents the attempt to develop this type of instructional outcome.

*Understandings.* Knowledge alone, embodied in the psychologically unsound truism that "knowledge is power," represents a much lower and less functional instructional outcome than that referred to in the psychologically sound statement that "understanding passeth knowledge."<sup>3</sup> Knowledge without power, or without understanding of its significance, is useless. Understandings are probably similar to but at a higher level than concepts. Understandings even more than

<sup>3</sup> Harl R. Douglass and Herbert F. Spitzer, "The Importance of Teaching for Understanding." *The Measurement of Understanding*, Forty-fifth Yearbook of the National Society for the Study of Education, Part I. University of Chicago Press, Chicago, 1946. p. 7.



concepts appear to be essential prerequisites to the functional use of what has been learned. Modern schools are increasingly stressing the development of understandings in pupils.

*Applications.* Abilities of pupils to apply the results of learning in logical thinking and in solving problems represent an end product or an ultimate goal of teaching. Skills, knowledges, concepts, and understandings all contribute to the attainment of this outcome. The development of realistic tastes and preferences is also prerequisite to the effective use of what has been learned in a functional situation. Logical thinking and problem-solving are not limited to mathematics, where the terms have perhaps most often been applied, but extend to any area, whether social, economic, political, or scientific, in which problems exist and decisions are to be made.

### Content of the informal objective test

It is highly important that the test be definitely based upon the objectives and outcomes of the course, and also upon the course content. It is true, naturally, that content is basic to a test, and furthermore that the best source of content material is found in the course itself. However, the measurement of factual knowledges, and the assumption that the pupil is necessarily able to use the knowledges he has acquired or been modified by, are unsound. Tyler found that knowledge of facts and ability to apply principles to new situations are related only to the degree shown by an average correlation coefficient of not much above .25 in science courses at Ohio State University.<sup>4</sup> Therefore, not only should the test be so constructed as to measure the degree of attainment of the pupils in the desired outcomes but it should do so by means of test situations that involve the ability to apply and use facts as well as knowledge of facts.

Care should be taken to sample course content widely and impartially in the selection of materials for a test. It is also ordinarily desirable to use more than one type of objective item in the test, but, on the other hand, not to use too great a variety of item types. For ordinary classroom tests given during one period, two or three types might be used; for longer examinations, variety might be increased

<sup>4</sup> Ralph W. Tyler, "Identification and Definition of the Objectives To Be Measured." *The Construction and Use of Achievement Examinations*. Houghton Mifflin Co., Boston, 1936. p. 7.

by using four or five types or modifications. It should be kept in mind that the subject matter itself is often a factor limiting the types of items used. Since recall items place a greater demand upon the pupil's memory of specific facts than is true of recognition items, it might well be expected that the pupil would recognize the accuracy of certain facts presented to him but not necessarily be able to recall the facts without clues. Therefore, recall items should be used only for important facts.

The test maker usually finds it advantageous first to construct items that fall into large groupings, such as matching exercises, and then to construct items having narrower scope. It is also desirable to construct multiple-choice items prior to alternate-response forms. This does not mean that all matching and multiple-choice items should be constructed before any true-false or simple recall items are made, but rather that first consideration should be given for a certain fact or relationship to the possibility of its use in an item form which is not so flexible and widely applicable as are the true-false and simple recall. If a particular idea does not, for example, readily combine with other similar relationships into a matching exercise and does not furnish enough plausible alternative responses for use in multiple-choice form, it might immediately be set up in one of the simpler forms.

The teacher will find it advantageous, for reasons that will be brought out clearly below, to write each item or each test unit on a filing card or slip. Alternate-response, multiple-choice, and simple recall items should be put on separate cards. Paragraph completion and matching exercises should be written on cards in their entirety, for such test units cannot be broken down by items for listing on separate cards. It is possible and desirable to code these cards in terms of the content they cover and also to keep records of the use of each item in the test and its validity. More will be said of these last two points in a later section of this chapter.

### **Assembling and preparing the informal objective test**

After the test items have been constructed, they should be sorted by types and carefully evaluated in their new settings. There should be a minimum number of items which all pupils can answer correctly or for which no pupils can get the correct answers. A difficulty level averaging about 50 per cent is recommended by Lindquist as most



satisfactory.<sup>5</sup> Items should therefore range from that point toward very hard and toward very easy. If there should be too few items of a certain type for a section of a test, those items should be re-drafted to fit into one of the sections definitely decided upon.

Test length depends on many factors other than the nature of the test items and the amount of time available for testing, but these are basic issues to be considered in the preparation of a test. The test should be of such length that all or very nearly all of the pupils can complete it before the end of the testing period. Recommendations have been made concerning the number of items of each type that can be given per unit of time at various age levels. From the available evidence it is impossible to determine in advance the exact working time required for a given form of objective examination. However, a reasonable estimate may be reached by allowing one minute of working time for each two recall items, each two multiple-choice items and each three true-false items. Such recommendations seem to have only very general significance, however, for the difficulty of the items and the age level of the pupils have much to do with time requirements, and teachers vary a great deal in the types of items they construct. The teacher will learn after brief experimentation how long a test should be for a given period of time. The number of items can be determined automatically by the number that have been constructed when the teacher considers the test to be complete and adequate and of proper length for the testing period. It is, however, important that a fairly large number of items be used in all objective tests.

Items should be arranged in parts or sections according to type in the final test. There is little agreement among test workers concerning the best arrangement of items for informal objective tests. Some prefer arrangement of items in each part by an increasing order of difficulty. If this method is used, the teacher's judgment concerning item difficulty is the only basis for arrangement when items are first used. Item-counting procedures furnish evidence on difficulty after items have been used with a class. Other persons prefer to arrange the items topically within each section of the test, and to consider item difficulty in the arrangement of items only by introducing the test by a few very easy items so that pupils will

<sup>5</sup> E. F. Lindquist, "The Theory of Test Construction." *The Construction and Use of Achievement Examinations*. Houghton Mifflin Co., Boston, 1936. p. 32-33.

not become discouraged before they get well started. The authors believe that either organization of the test is satisfactory and that the individual teacher should use the method which adequately meets the conditions under which he uses the informal objective examination.

The examination should be prepared for use with the pupils by a mimeographing or other method of reproduction if possible. Some item types can be given orally if absolutely necessary, and the black-board can be used for short quizzes. Complete directions to the pupils should always be provided. This sometimes entails general instructions at the beginning and separate directions for each part of the test. If the item forms are difficult to understand or if pupils are taking objective tests for the first time, samples showing how they are to record their answers should be given with the directions. The samples should be so simple in content that they will be readily comprehended by all pupils. Illustrations of directions to pupils and of samples to demonstrate methods of answering test items are given later in this chapter.

Pupils should be told in the directions whether or not to guess, and should also be told how the test will be scored. The most common procedures and those usually recommended are to instruct the pupils not to guess and then to correct their scores for guessing on alternate-response items. On the other hand, pupils are usually told to attempt each item on the matching test.

### **Administering and scoring the informal objective test**

Little need be said here concerning the administration of the informal objective test except to point out that if the directions to pupils and any necessary sample items are carefully and well prepared the actual administration of the test is simple indeed. The teacher should be careful not to give intentional or unintentional assistance to individual pupils by answering any questions they may ask. The safest procedure is to make certain that the pupils understand how to take the test by careful preparation of the directions, to make sure that individual test items require no explanations by framing them with care, and then to answer no questions about word meanings or interpretations to be placed on certain items while the test is in progress. Pupil questions concerning typographical errors they may encounter in the test should be investigated and the at-



tention of the entire class should be called to any such errors that might within reason cause misinterpretations of items.

Scoring of the test should be by the predetermined method, and should vary with the type of objective item. Scoring keys can be prepared easily by using a copy of the test and cutting it into strip keys and cutout stencils as required. With such keys available, the actual mechanics of scoring the tests are very simple. Each correct answer should ordinarily be given one point of credit. It will be advantageous to mark each correct answer with a colored pencil for later use for instructional purposes.

Chances of guessing the correct answers vary with different item forms. There is little if any chance of guessing, or at least of making a pure guess, on recall item forms. Obviously, the chance is 50-50 on an alternate-response item, but it is only one in five for a multiple-choice item with five alternatives. The correction for chance formula is

$$\text{Score} = \text{Rights} - \frac{\text{Wrongs}}{N - 1}, \text{ or } R - \frac{W}{N - 1},$$

where  $N$  represents the number of possible answers to an item. For the true-false item, this becomes  $R - W$ . For multiple-choice items of 3, 4, and 5 alternatives, the formula becomes respectively

$$R - \frac{W}{2}, R - \frac{W}{3}, \text{ and } R - \frac{W}{4}.$$

Correction for chance is ordinarily used with the true-false test and the multiple-choice test consisting of items that have as few as three alternatives. It need not necessarily be used with multiple-choice items having four or more alternatives, as the chance of making a correct guess is not great in such tests. Matching tests are not corrected for chance, for little opportunity for guessing exists if they are properly constructed.

There should be no attempt to weight individual items of a test differently according to their importance or difficulty. A summary of various studies dealing with this question leads to that conclusion.<sup>6</sup>

<sup>6</sup> J. Murray Lee and Percival M. Symonds, "New-Type or Objective Tests: A Summary of Recent Investigations (October 1931-October 1933)." *Journal of Educational Psychology*, 25:161-84; March 1934.

It may be desirable in some instances, however, to assign varying weights to the scores resulting from different parts of the test in order to account for differences in difficulty or average time required per item, in which case the most satisfactory procedure is probably to multiply by 2 or by 3 the scores from test parts that are thought to be deserving of extra weighting.

### Anticipating future testing needs

For the teacher who repeats courses annually or more than once each year, concern with a particular informal objective test should not end with the final direct use of the results. Informal objective testing is not economical of teacher time if the teacher starts afresh in the construction of every test over a period of years. Construction of informal objective tests should be a cumulative and selective process resulting in constant improvement of the tests actually used in the classroom. If tests are to be evaluated and improved in the manner suggested below, test booklets should not be returned to the pupils permanently. However, they may well be distributed for review purposes after the test has been scored, and collected when the instructional purpose has been accomplished, or used with individual pupils in conferences concerning special points needing further emphasis in their work.

As a means of determining the validities of individual items for future use, the teacher will find the method generally known as item-counting of great value. One of the simple item-counting methods is based on a division of the class into groups of above-average and below-average performance on the test, with about half of the class in each group. The test papers should then be sorted into corresponding groups. The number of correct responses to each test item by the pupils in each group can then be determined by a routine clerical procedure. This ordinarily involves the use of squared paper on which the columns represent the items of the test and the rows are used for checking the items correctly answered by each pupil. A summation of the check marks in each column for each of the two pupil groups is then made. When the number of correct responses to each item is converted into a percentage of the number of pupils in the group, data essentially of the type shown in Table 2 of Chapter 5 become available.



Such evidence is valuable to the teacher in determining which test items properly discriminate pupil abilities by showing higher percentages of correct answers for above-average than for below-average pupils, which ones might be suspected of ambiguity or other faults because of failure to effect such discrimination, and which, if any, show reversals of the desired type of discriminative power. If the information concerning item validities thus obtained is recorded on the cards which it was suggested in an above section should be set up for test items and groups of items, the cards become a valuable file for use in the construction of future tests. Items that show the proper type of discrimination can be retained, and those that discriminate in the wrong direction can be discarded or revised after critical examination reveals the source of their ambiguity or other weakness. Ultimately, the card file should include only test items that have been found satisfactory in actual classroom measurement.

A card file of this type can be used for the construction of new tests when the occasion arises, with assurance that the ambiguous items occurring in previous tests have largely been eliminated. It is, of course, desirable to add to the file as course content changes and to withdraw items which, although valid, are no longer applicable because of changing course content and objectives. Need for such constant turnover is greater in the social studies and sciences, in which current developments perhaps have the greatest immediate influence, than in subjects for which the content changes less rapidly, but it is undesirable for any course that objective classroom testing be allowed to become static.

Although this procedure for validating test content may on the surface appear to be lengthy and somewhat involved, the teacher will realize significant dividends in improved pupil measurement by the use of it or some similar procedure. After such a system of keeping a cumulative test item file is once established, the teacher will realize the great saving in time and the increased testing efficiency that results. Time expenditure by the teacher is greatest for the typical essay test in the scoring of pupil results. Time expenditure by the teacher is greatest for the informal objective test in its preparation. Attention to the construction of good tests seems much more defensible than attention to the scoring of tests which in many instances are not satisfactory measurement instruments.

## 4 TYPES OF OBJECTIVE ITEMS

The uses and limitations of the five basic types of objective items are discussed below and a few sample items<sup>7</sup> are given to illustrate each type.

### Simple recall items

Simple recall test items cannot be definitely distinguished from completion exercises. The major distinctions appear to rest on complexity and length of the test unit and perhaps on the number of pupil responses called for. The simple recall form is by far the most widely used of the recall item types. It usually involves a very brief response by the pupil, such as writing a word, a number, a symbol, or a short phrase in a designated place in answer to a question or to complete a statement.

*Uses and limitations of simple recall items.* The simple recall item is best adapted to the measurement of rather highly factual knowledges of the *who, what, when, where* types, and is very widely adaptable to different subject matter in such uses. It can be used to test the ability to identify things described or pictured, in which form it has rather wide range. In identification exercises, it is perhaps best adapted for use with maps and charts in the social studies and representations of biological structures in the natural sciences. It is useful in computational problem situations in arithmetic and the physical sciences.

One of the major characteristics of the simple recall form is its apparent ease of construction, which tends to encourage wider use than is perhaps justified. Because of its tendency to measure factual knowledges rather than understandings, there is danger of overweighting tests with factual materials if the simple recall item is too widely employed. This item is not readily adaptable to the measurement of abilities to apply facts, to perceive complex relationships, and to draw logical inferences. The simple recall form is readily understood by pupils because of its similarity to the essay question.

<sup>7</sup> For extensive samples of major item types and their modifications classified by instructional outcomes and item types, see J. Raymond Gerberich, *A Guide to Achievement Test Construction: Specimen Objective Item Types*. Longmans, Green and Co., New York, 1953.



The simple recall item is difficult to score because of the tendency for the responses to lack complete objectivity, even though responses may be provided for in terminal and aligned form. It is further limited by the fact that it is not directly adaptable to machine methods of scoring.

*Major types of simple recall items.* The simple recall item is perhaps most frequently presented in the form of a declarative statement with a blank in which the pupil is to write the correct completion occurring at the end of the sentence. It also is frequently used, particularly in the lower grades, in the form of a question to be answered by the pupil on the line immediately following. Another form less widely used but satisfactory involves a list of terms or statements introduced by directions which tell the pupil to write on the line following each the other term or statement called for by the directions.

### Excerpt from National Achievement American History Test <sup>8</sup>

#### READ THE FOLLOWING SAMPLE:

The inventor of wireless was Marconi.

In this SAMPLE, the name "Marconi" was written on the line to finish the sentence.

*DIRECTIONS: Finish every sentence by putting the correct name on the line.*

#### PRACTICE EXERCISE:

The first President of the United States  
was \_\_\_\_\_

- |   |  |
|---|--|
| 1. The Rough Riders were under the leadership of a man named _____                                | 4. The builder of the Panama Canal was _____ Major-General _____     |
| 2. A reaper was invented in 1834 by a man named _____ Cyrus H. _____                              | 5. Many railroads were developed by a man named _____ James J. _____ |
| 3. In the campaign of 1896, the unlimited coinage of silver was favored by _____ William J. _____ | 6. The North Pole was discovered by _____ Robert E. _____            |

## Completion items

Completion items may be either of the sentence or the paragraph type. Frequently there is little by which a sentence completion item can be distinguished from the simple recall item. The more typical form of the completion exercise, however, is that based on a paragraph of unified material in which several blanks are provided for the pupil to fill with the words, numbers, or short phrases that correctly complete the meaning. Since blanks in the completion exercise

<sup>8</sup> Robert K. Speer, Lester D. Crow, and Samuel Smith, *National Achievement Tests: American History*, Grades 7 and 8. Published by Acorn Publishing Co., 1939.

only occasionally occur at the ends of sentences, pupil responses typically are scattered over the page. An adaptation of this type of exercise places a number in each blank and similarly numbered blanks at the right-hand margin for use by the pupils in recording their answers. This results in simplifying the scoring procedure for completion exercises.

*Uses and limitations of completion items.* Similarities between the simple recall item and the sentence and paragraph completion exercise result in considerable similarity of their uses and limitations. Both are typically rather highly factual, but the latter requires the pupil to handle a larger unit of thought and to integrate his ideas more fully. Both are difficult to score objectively, and must be so constructed that the blanks call for definite responses. Neither can be scored directly by mechanical methods. Both may become puzzle situations for the pupil if too much of the thought is omitted from the statement to permit of reasonably quick comprehension of meaning by the pupil. The completion exercise is somewhat harder to score than the simple recall item unless a device that results in aligned and marginal responses is employed.

Completion examples are not so widely adaptable as simple recall items because of the need for broader and more unified thought units in the former. However, both forms are useful with a wide variety of content. The completion sentence is applicable, for example, in situations involving use of the correct language form in a given setting in English or the foreign languages, in completing arithmetical examples of the equation form, and in a variety of situations in the social studies and sciences. The paragraph completion exercise is useful in various courses for situations in which a chronological, organizational, sequential, or cause and effect type of pattern exists, as, for example, with the processes involved in a complete cycle of blood circulation in the human body.

*Major types of completion items.* Sentence completion exercises frequently require the filling of two or more blanks by the pupil and the blanks do not, of course, occur at the ends of the sentences, as they typically do in simple recall items. The paragraph completion exercise differs from the sentence completion mainly by consisting of a longer and perhaps more complex thought unit, probably by requiring more pupil responses, and by consisting of two or more sentences in a well-unified paragraph.



Excerpt from Metropolitan Reading Test <sup>9</sup>

**DIRECTIONS.** In each paragraph a blank line means that a word has been left out. Read each paragraph. Then think of the word that should be in each blank. Write the word in the parentheses at the side of the page. You should get the answer from the paragraph itself.

**SAMPLE.** Dick, Tom, and Fred are brothers. The names(\_\_\_\_\_) <sup>a</sup>  
of Dick's brothers are (a) and (b) . . . . (\_\_\_\_\_) <sup>b</sup>

24-27. Many bad automobile accidents happen as a result of drivers' going to sleep at the wheel. If a driver feels (24), he should consider it a danger signal. (\_\_\_\_\_) <sup>24</sup>  
There are many causes of (25). Sometimes the hum (\_\_\_\_\_) <sup>25</sup>  
of the motor or the rapid passing of objects makes a driver sleepy. Some drivers get drowsy at certain times of day. Loss of sleep causes drowsiness, too. (\_\_\_\_\_) <sup>26</sup>  
Some drivers pull off the (26), stop the (27), and take a nap when they are sleepy. . . . . (\_\_\_\_\_) <sup>27</sup>

## Alternate-response items

Alternate-response items are those in which only two alternatives are presented to the pupil for his response. The simplest and most common forms of alternate-response items are the true-false, requiring an answer concerning the truth or falsity of a statement, and the yes-no, requiring one of those answers to a question. Another form involves the selection of the correct one or better one of two alternatives that are presented as possible completions in a given setting.

The true-false, as the most widely used alternate-response type, has doubtless been the most popular form of recognition item and probably remains so today for classroom testing purposes. It typically involves a very simple method of response by the pupil in aligned answer positions at either the left or right side of the test paper.

*Uses and limitations of alternate-response items.* The true-false item is widely applicable in all subject fields. Its ease of construction has resulted in greater popularity and wider use than have been attained by any other item form. However, its ease of construction is frequently delusive, for the elimination of ambiguities from the true-

<sup>9</sup> Richard D. Allen and others, *Metropolitan Achievement Tests, Test 1, Reading, Advanced Battery*. Published by World Book Co., 1946.

false item is sometimes difficult to accomplish. Since this weakness seems to be inherent in the item itself, test technicians are tending to use it less and less. It and the simple recall item are perhaps most frequently taken almost verbatim from textbooks, and consequently in such cases a premium is placed upon photographic memory for facts by pupils.

Alternate-response item forms have the advantage of affording coverage of many individual items in a short period of time, since the time requirements are less than for most item types. On the other hand, guessing is more of a problem for this than for any other item type, for which reason little diagnostic value can be obtained by using an item-count method of analyzing the results for a group of pupils or an individual pupil. Alternate-response items are highly objective in scoring, and are readily understood by pupils. This item type is readily scorable by mechanical methods in all of its common varieties.

True-false items can be used satisfactorily in many situations if they are constructed carefully enough to keep them free from ambiguity. They are especially useful for situations in which the absence of enough plausible alternative responses makes the use of a multiple-choice item impracticable.

The type of alternate-response form that requires the pupil to select the one of the two alternatives that correctly fills a particular need is very widely useful for measurement of a functional type of instructional outcome in English and the foreign languages. It could be used in a wide variety of situations, but in practice this item form has been limited largely to language usage testing.

*Major types of alternate-response items.* The most common form of true-false item may be set up so that the pupil will respond by encircling or underlining a *T* or *F*, or a *True* or *False*. The arrangement of answer spaces in columns under *T* and *F* in which the answer is indicated by an "X" or check mark has the added advantages of speed of response and ease of scoring.

Another common form is presented as a question, the pupil's responses usually consisting of encircling or underlining either *Yes* or *No*. This form, which differs little from that presented above, is preferable for use with young children because the situation presented is a very normal one. An alternate-response form commonly used in English and foreign language tests involves the selection of the proper one of two given word forms for use in a certain setting and indica-



tion of the one selected by crossing out the incorrect word form or marking the correct word form.

Excerpt from *Progressive Tests in Related Sciences*<sup>10</sup>

DIRECTIONS: Read each statement below. If the statement is TRUE, you are to mark the letter T; if it is FALSE, mark the letter F.

41. Water power is used to produce electrical power. T F<sup>41</sup>
42. Most of our paper is made from wood. T F<sup>42</sup>
43. The dog was one of the first animals to be tamed by man. T F<sup>43</sup>

### Multiple-choice items

Multiple-choice items have come to be the most popular form for standardized testing of recent years, and are increasingly coming into wide use for informal objective testing as well. A recognition item type, the multiple-choice item commonly consists of an incomplete statement followed by from three to five responses that will complete the statement with varying degrees of accuracy. The pupil is expected to choose the response that correctly or best completes the statement, and typically to indicate his choice by an answer appearing in a column at the left or the right side of the test paper.

This item type may be in question rather than in statement form or may consist of three to five words, symbols, or numbers from which the correct one is to be chosen by the pupil. It may request the best of several correct or partially-correct answers on a given point. It may even require responses for the two or more correct answers among those furnished, in which case it becomes a multiple-response item.

*Uses and limitations of multiple-choice items.* The multiple-choice and its numerous variants perhaps represent the most valuable and

<sup>10</sup> Georgia S. Adams and John A. Sexson, *Progressive Tests in Social and Related Sciences, Test 6, Elementary Science*, Elementary Battery. Published by California Test Bureau, 1946.

at the same time the most widely applicable type of objective test item. It is readily, although not necessarily easily, adaptable to the measurement of discriminative power, inferential reasoning, interpretive ability, reasoned understanding, generalizing ability, and other types of outcomes deriving from the pupil's ability to apply and use facts. It is not difficult for pupils to understand and use. It is highly objective, and can be readily scored either by hand or by machine. Item-counting procedures based on the results for an individual pupil or a class have considerable diagnostic and analytic significance.

Multiple-choice and multiple-response items in their variety of forms are so widely adaptable to different types of content that the preceding discussion should make the fact evident without illustration. As is the case for the true-false item, there is probably no field of learning to which the multiple-choice item is not widely applicable. However, the necessity for finding at least two and in many cases as many as four plausible responses to go with the correct completion somewhat limits the applicability of the item form within each subject field. Ingenuity on the part of the test maker and the results of practice in item construction make the item type very widely applicable to the content of various instructional areas, however. Multiple-choice items are not as easily constructed as are some other objective test forms, for there are various technical problems that require great care in the drafting of items. The incorrect answers pupils give to simple recall items often serve as excellent incorrect alternatives if the item is converted to multiple-choice form.

*Major types of multiple-choice items.* The basic and probably most common multiple-choice form is that in which the correct or best completion is to be selected by the pupil from the three to five that are furnished for an incomplete declarative sentence or in answer to a question.

A common use of multiple-choice forms is in testing various types of reading ability, as, for example, ability to comprehend the meaning of a paragraph, by basing a single item or several items on a passage of reading material in English or a foreign language. Somewhat similarly, multiple-choice items can singly or by groups be based on a map, chart, diagram, or table, and require the pupil to interpret the data presented as a basis for answering.

Another variation, called the multiple-response, is that in which



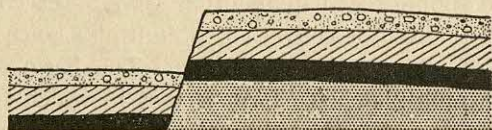
the pupil is asked to select all of the correct completions from the three to five typically given. There may be only one or as many as several correct answers to each item when this form is used. Each

Excerpt from Cooperative Social Studies Test <sup>11</sup>

3. The period of change from hand-made to machine-made goods is known as the
  - 3-1 Industrial Revolution.
  - 3-2 Handicraft Age.
  - 3-3 Age of Big Business.
  - 3-4 Reformation.
  - 3-5 Renaissance. .... 3(    )
  
4. Which of these was *not* one of the original thirteen states?
  - 4-1 Virginia.
  - 4-2 Georgia.
  - 4-3 Massachusetts.
  - 4-4 Florida.
  - 4-5 New York. .... 4(    )

correct response is ordinarily assigned one scoring point of credit. The fact that not only the choice but also the response is plural accounts for the distinction in names between this and the more common multiple-choice item.

Excerpt from Read General Science Test <sup>12</sup>



28. The geological formation above constitutes evidence of —
  6. volcanic action.
  7. erosion.
  8. folding.
  9. sedimentation in a running stream.
  10. movement in the earth's crust.

<sup>11</sup> Harry D. Berg and Elaine Forsyth, *Cooperative Social Studies Test for Grades 7, 8, and 9*, Form X. Published by Cooperative Test Service, 1947.

<sup>12</sup> John G. Read, *Read General Science Test*, Form A. Published by World Book Co., 1950.

## Matching exercises

Matching exercises are in effect combinations of multiple-choice items in such a manner that the choices are compound in number. Matching exercises differ from all of the objective forms treated previously in the fact that they must occur in groups. There is really no such thing as a matching test item, unless a correct pairing pulled from a group of which it is a part might be so designated. Matching tests are by nature, then, multiple in type, and the number of scoring points is ordinarily determined by the number of responses required of the pupil.

A matching exercise or set usually consists of two lists of related facts between which a constant type of relationship exists throughout. The pupil's responses are expected so to pair items in the two lists as to indicate their proper relationships. Variations involve unbalanced sets, in which more items occur on one side than on the other, sets in which items of one side may be used more than once each, and even compound sets in which double or even triple matchings of all items are necessitated by the provision of three or even four related lists instead of the customary two.

Pupil responses to matching exercises are usually in the form of identifying numbers or letters written in column form in parallel with the items in one of the two or more lists. The unbalanced set has the definite advantage of reducing the chances of guessing the correct answers to practically zero.

*Uses and limitations of matching exercises.* Matching exercises are likely to be rather highly factual in nature, and to make use of the *who, what, when* and *where* types of relationships and of identifying or naming abilities. They are rather easy to construct, and are perhaps for that reason more widely used than their characteristics warrant. They are likely to include clues to the correct responses unless there is rigid adherence to uniform categories of items in a matching set, and this restriction, desirable though it is, limits at least one side of the test unit to numbers, words, or at least short phrases. This restriction in turn tends to limit use of the item form mainly to factual types of subject matter.

The matching exercise is economical of space and of construction time. It is useful for matching terms and definitions, names and events, events and dates, books and authors, causes and effects, generalizations and applications, words and symbols, English and



foreign words, and many other pairs of related items by use of verbal lists. It is also useful with numbered maps, charts, or pictorial representations for matching places and names, places and events, trends and dates, or objects and names in great variety. The matching exercise appears to be most useful with factual knowledges in a great variety of situations where it is desirable to test over a number of comparable relationships.

*Major types of matching exercises.* The fundamental form of matching exercise has an equal number of items in both lists and involves the use of all of the items in the pairing. Unbalanced matching sets provide more items on one than on the other side and require that only as many of the items of the longer list be used as have proper pairings with the items of the shorter list.

#### Excerpt from Metropolitan Literature Test <sup>13</sup>

**DIRECTIONS.** In the parentheses after each character in Column 2 put the number of the character from Column 1 that appears in the same story or poem.

COLUMN 1	COLUMN 2
1. Ichabod Crane	53. Marygold.....( ) 53
2. Don Quixote	54. The Mayor of Hamelin ..( ) 54
3. Laurie	55. Jim Hawkins.....( ) 55
4. The Pied Piper	56. Nello.....( ) 56
5. Midas	57. Dulcinea.....( ) 57
6. Aunt Harriet	58. Brom Bones.....( ) 58
7. Patrasch	59. Elaine.....( ) 59
8. Black Beauty	60. Elizabeth Ann.....( ) 60
9. Launcelot	
10. Long John Silver	

Diagrams, maps, charts, and pictures may be used in what are often called identification exercises by requesting the pupil to match identifying names of places, objects, or parts with their representations in the accompanying figure or picture.

## 5 CONSTRUCTING OBJECTIVE TEST ITEMS

This section of the chapter considers the general principles to be followed in the construction of various objective item types. Such questions as adaptation of item types to various subject matter

<sup>13</sup> Richard D. Allen and others, *Metropolitan Achievement Tests, Test 6, Literature*, Advanced Battery. Published by World Book Co., 1946.

and the construction and use of the test as a whole were considered earlier in the chapter. Because of the multiplicity of item types, it is impossible to discuss all of them in detail. Therefore, the suggestions are intended mainly for the basic or most common forms of items, although in many instances they are equally well adapted to modified types of the basic items.

General suggestions that seem to be equally applicable to all objective item types are given in the following section. These should serve as the introductory portion of the lists of suggestions on later pages for the various common or basic forms of items. The student will find that frequent reference to the sample items in the preceding section will be helpful in the study of methods for constructing the various item types. It should be apparent that common sense and personal experience must furnish the basis for recommendations on many issues discussed. Objective evidence is not available concerning the relative merits of different approaches on many of the issues, and on other points only inconclusive evidence and conflicting opinions and practices are presented in the educational literature. Therefore, this section can be said to present the authors' views, based on objective evidence and opinions of others and on their own experience in test construction, on a variety of detailed points which must be considered if objective item types are to be well constructed.<sup>14</sup>

### General suggestions for constructing objective items

A number of the suggestions given here apply equally well to all or most objective item types. Attention will be given in the subsequent pages to suggestions that apply to recall types and to specific item types of the recognition form.

(1) *Rules governing good language expression should be observed.* This point deserves mention because carelessly framed and ungrammatical items are more likely to be subject to misinterpretation than are items that are carefully constructed and correctly stated.

(2) *Difficult words should be avoided.* Care should be taken at all times to make certain that the words used in objective items are known to all pupils, for every pupil should be able to understand the intent of all items. This recommendation does not, of course,

<sup>14</sup> See also Robert L. Ebel, "Writing the Test Item." *Educational Measurement*. American Council on Education, Washington, D. C., 1951. p. 213-44.



apply to the technical words of the subject being tested, for knowledge of technical vocabulary is an outcome of instruction that may well be tested. Every effort should be made to adapt general vocabulary words to the ability levels of the pupils being tested, however. In case of doubt, it is always a safe procedure to choose the simpler of two words that might be used in stating a test item.

(3) *Textbook wording should be avoided.* It is undesirable to obtain items merely by taking a statement from a textbook and using it in its exact textbook form or with a negative inserted, a word omitted, or other minor adaptation. In the first place, an occasional pupil has a memory for specifics of what he has read or heard which would enable him to answer the item from memory rather than in terms of knowledge and understanding. In the second place, a majority of textbook sentences, unless they are from summary paragraphs or are topic sentences of paragraphs, are too detailed to merit direct attention in a test. In the third place, there is danger that items so selected would be too much dependent upon a particular textbook or author and not be broadly representative of the field being tested.

(4) *Ambiguities should be avoided.* Care should be taken to make certain that each test item is subject to one and only one interpretation. It is not always easy to accomplish this purpose, for ambiguities sometimes remain after an item has been carefully framed and scrutinized. Items should be sufficiently definite that there is no chance for misinterpretation of meaning through reasonable implications or logical inferences. Item-counting methods of evaluating items after they have been used once are helpful in eliminating ambiguities that have been overlooked in the initial framing of a test.

(5) *Items having obvious answers should not be used.* Items to which answers are obvious have no value in a test and should definitely be avoided.

(6) *Clues and suggestions should be avoided.* Items containing clues or suggestions also contribute nothing to a test and may well lack validity.

(7) *Items that can be answered by intelligence alone should not be included.* Items that depend not at all on knowledge or understanding but that can be answered by the exercise of intelligence alone have no place in an achievement test.

(8) *Quantitative rather than qualitative words should be used.* It is preferable to use words that have quantitative and definite

meaning rather than words that are qualitative in nature, as a means of eliminating items that depend on opinion rather than upon facts.

(9) *Catch words should not be employed.* There is no justification for the inclusion in achievement test items of catch words, misleading statements, or irrelevant confusions. Pupils recognizing such points might interpret such features as typographical errors or as unintentional for other reasons and answer them in terms of what they thought was intended. Furthermore, the best readers, who are frequently the best pupils, are perhaps least likely to note minor errors in a test because rapid reading entails less attention to specific letters and even words than does slow reading.

(10) *Items should not be interrelated.* Items should not ordinarily be so related, at least if they are adjacent or close together in the test, that one depends on one or more other items in such manner that an answer to the first determines responses for the related item or items. In effect, such dependence places more than the intended amount of weight on the first item of any such sequence when answers consistent with the first are given by a pupil for subsequent items.

(11) *Response positions should preferably be aligned.* It is preferable, although not always possible, to have the response positions occur in a columnar arrangement. The pupil is aided by such a consistent position for responses and scoring of the results is greatly facilitated.

## Suggestions for constructing recall-type items

Several suggestions applicable to recall item types alone are given and briefly discussed here. These suggestions represent for recall items a continuation of the list of general suggestions in the preceding pages of this chapter. As the simple recall and completion types are very similar except for two of the following points, the recommendations for these item types are included in one list.

(1) *Lines for responses should be of the same and of adequate length.* In recall item forms the length of all lines or blanks provided for pupil responses should be the same. The lines or blanks should be long enough to provide for normal writing of the longest word likely to be given as an answer. The constant length of line avoids giving any clue to the length of the correct answer that might be of use to the pupil in choosing between two answers he might be considering.



(2) *Desired responses should be definite.* Each recall item should require a definite idea or concept as the correct answer in order to reduce the possibility of misunderstanding by the pupil and to insure objectivity of scoring. The response may be a word, a date, a number, a symbol, a formula, an answer to a problem, or even a short phrase.

(3) *Desired responses should be important.* Only important and crucial aspects of a statement should be omitted in recall forms of items, for the omission of secondarily important or unimportant aspects of a statement reduces the significance of the item.

(4) *Any correct answer should receive credit.* Any answer that is correct, whether or not it is the one the teacher expected, should receive credit and the answer should be added to the scoring key for future use.

(5) *Spelling errors probably should not be penalized.* Unless spelling errors occurring in pupil answers are in words technical to the subject for which the test is given, scoring should probably be in terms of the pupil's intent rather than in terms of his spelling accuracy.

(6) *"A" or "an" should not immediately precede a blank.* Either of the indefinite articles restricts the nature of the response word to follow in terms of grammatical correctness, so that the range of possible correct answers is mechanically narrowed for the pupil when "a" or "an" immediately precedes a response position. Employment either of the definite article "the" or of "a(n)," which means either "a" or "an," is permissible.

(7) *Positions for responses should ordinarily be at the ends of the sentences.* It is preferable that blanks to be filled occur at the end rather than in the middle of sentences. Statements can usually be so worded that this is easily accomplished.

(8) *Completion paragraphs should be unified wholes.* A completion paragraph should be unified and well organized and should not consist of several unrelated or poorly related sentences. The pupil's ability to grasp the entire thought unit should be essential to correct responses for the several blanks in the paragraph.

(9) *Completion paragraphs should not obscure the meaning by containing too many blanks.* Sufficient of the paragraph should be given that the meaning is clear to an informed and intelligent reader. It is easy for the teacher constructing a paragraph, who knows definitely what the paragraph is about, to assume unconsciously that the pupil should have the same knowledge and consequently leave

out so many words that to the pupil the meaning is obscure or not ascertainable.

### Suggestions for constructing alternate-response items

The suggestions below for the alternate-response type of item supplement the general suggestions previously discussed. As the true-false is the most widely used of these types, most of the suggestions below relate primarily to it or a closely allied form.

(1) *Double negative statements should be avoided.* Double negatives serve no useful purpose, but they may cause needless and harmful reading problems for some pupils.

(2) *Statements that are part true and part false should not be used.* Statements should be either true or false, for the use of a true major clause and a false dependent clause or of some other combination of truth and falsity is confusing to the pupil and adds nothing to the test. Although such part true, part false statements are used by some test workers, the result frequently is an unintentional "catch" item.

(3) *"Specific determiners" should be used sparingly and carefully.* Such specific determiners as "always" and "never" occur in false statements much more frequently than in true statements. Statements containing cause or reason clauses also tend to be false more often than true. On the other hand, comparison statements and very long statements are more often true than false.

(4) *Answers should be required in a highly objective form.* It is inadvisable to have pupils write a letter, such as *T* or *F*, or a word, such as *True* or *False*, in answering the items, for those letters and words look much alike when poorly written or when written with the attempt to confuse the scorer. Methods requiring pupils to encircle or to underline *T* or *F*, *Yes* or *No*, or having pupils mark an "X" in the brackets in either the *T* or *F* column are to be preferred.

(5) *Approximately an equal number of true and false statements should be used.* It is not desirable to have a great imbalance of true and false statements, but on the other hand there is no need for exactly the same number of each type of item.

(6) *Random occurrence of true and false statements should be employed.* A coin may be tossed or some other simple chance procedure be used to make certain that true and false statements occur in random or chance order.



## Suggestions for constructing multiple-choice items

The following suggestions, supplementing the general recommendations given in an earlier section of this chapter, are primarily for the multiple-choice item type with only one correct answer or the closely related best-answer type.

(1) *As much of the statement as possible should occur in the introductory portion or stem.* There is no justification for repetition of the same introductory word or words in each of the alternatives; the introductory, or common, portion of the item should include as much as possible as a means of saving space.

(2) *Alternative answers should all be stated in correct grammatical style.* It should be possible to follow the stem of an item with any one of the alternative answers and have the statement be grammatically correct.

(3) *Incorrect alternatives, or confusions, should be plausible.* One or more alternatives that are obviously incorrect in effect give the pupil a greater chance of guessing the correct answer. Pupils' wrong answers to recall items often provide excellent confusions for the same items if put into multiple-choice form.

(4) *"A" or "an" should not ordinarily be used to introduce the alternative answers.* Unless all answers can follow the same article with grammatical correctness, the "a(n)" device mentioned above or the indefinite article should be used to introduce the alternative answers.

(5) *Items should ordinarily have four or five alternative answers.* Except for use with very young children, four or five alternative answers are preferable as a means of reducing the chances of guessing the correct answer and in order to obtain the desired degree of item difficulty, although two well-chosen confusions are preferable to three or four implausible wrong answers.

(6) *All items should ordinarily have the same number of alternate answers.* Four- and five-response items should ordinarily not be mixed in the same test, for the same number of alternatives for each item is preferable for ease in correction for guessing.

(7) *Alternative answers should ordinarily occur at the end of the statement.* Although the responses may be so placed that additional material common to all is necessary to complete the statement, rewording will ordinarily make possible their placement at the conclusion of the statement.

(8) *Answers should be required in a highly objective form.* It is perhaps preferable that a pupil write the identifying letter or number for the intended response or encircle or otherwise mark it in a special answer column. There is little efficiency in a method requiring underlining or, worse yet, both underlining and otherwise indicating, an intended answer.

(9) *Correct responses should be distributed with approximate equality among possible answer positions.* In four-response items, for example, the first, second, third, and fourth alternatives should be correct for approximately the same number of items. It may be desirable to favor the centrally-located responses slightly over first and last responses for the correct answers.

(10) *Random occurrence of correct responses should be employed.* A die may be tossed (disregarding the six) or some other simple chance procedure be used to insure random order in the occurrence of the various correct answer positions.

### Suggestions for constructing matching exercises

The suggestions given below for the common type of matching set supplement the general suggestions on pages 187 to 189 for all types of objective items.

(1) *Only one correct matching for each item should be possible.* If items are not mutually exclusive, i.e., subject to only one correct matching, some pupils may be penalized because they happen to choose the one of two or more possible matchings for a certain item that results in the lack of a proper answer for an item at the end of the matching process, when the same number of items appears in each column.

(2) *Consistency of grammatical form should be used.* All items in the left-hand set should agree in form and all items in the right-hand set should likewise be in agreement. It should be possible insofar as the form of the statements is concerned to associate any item of the left with any item of the right column. If this is not true, answers can be obtained partly by attention of the pupil to grammatical detail in the statements.

(3) *Consistency of classifications should be maintained.* Each of the two lists should contain items that are of the same category. Although matching sets that are not consistent within each column are used by some test makers, the results from mixed categories are



sometimes confusing, often provide a means of answering items by the exercise of general intelligence alone, and in general are unsatisfactory. Consistent categories are much to be preferred.

(4) *Matching sets should neither be too long nor too short.* From ten to fifteen pairings are probably optimum for balanced-matching groups. More than fifteen pairs become cumbersome and time-consuming. Fewer than ten pairings present opportunities for good guessing on the last few matchings by the pupil who knows most of the pairings. Unbalanced matching sets are definitely preferable and perhaps should be used in all matching sets.

(5) *Items should be listed in random order in each list.* Such logical arrangements as alphabetical order of first letters of words and chronological order of dates usually accomplish this purpose, for such arrangements are not likely to have any similarity to the relationships between the items of the two lists and furnish no clues to the pupils.

(6) *A set of matching items should always be complete on one page.* The necessity for frequent rereading of items makes very inefficient any separation of a set of matching items by having it appear on two pages of the test.

(7) *Answers should be required in a highly objective form.* Perhaps the most satisfactory method of providing for pupil responses is to accompany one list with letters or numbers identifying each item and the other list by answer positions, and then to have pupils write the letters or numbers in the answer column in such manner as to indicate their choices.

## 6 USING RESULTS OF INFORMAL OBJECTIVE TESTING

Only brief mention is made here of the uses to which the informal objective examination can be put. The alertness and ingenuity of the teacher largely determine the values that result from his use of the informal objective test.

### Informal objective tests in instruction

The evaluation of pupil and class achievement is most effectively accomplished through the use of the objective examination. Even if there were standardized tests for the measurement of most of the outcomes of class instruction, they would be unsuited for this type

of use. Properly constructed objective examinations within certain limits aid the teacher in determining points at which instructional adjustments should be made. Pupils, likewise, may be led to discover their specific weaknesses in achievement. Informal objective test results are thus shown to have general diagnostic value for relative pupil strengths and weaknesses. Such tests can also be used for instructional as well as for measurement purposes. Informal objective drill and remedial devices can be constructed by the alert teacher.

### **Informal objective tests in determining course marks**

Pupils' scores from valid and reliable objective examinations afford the teacher's best single basis for measuring and rating pupil achievement within a given subject. The results of objective examinations enable the teacher to improve the reliability of his marks if the tests themselves are valid and reliable measures of the course outcomes. Teachers can learn with practice to construct course examinations that will satisfy the criteria of a good examination, and that will be more valid tests for the outcomes of his particular course than standardized tests could ever be. The remaining step for the use of test results in marking is to convert scores to the particular type of marks desired. Because of the importance of this use of informal objective test scores, a widely used method of converting them to course marks is explained in Chapter 13. This system can readily be adapted as required, if it is not applicable in its present form, to the marking system used in a particular school.

### **Topics for Discussion**

1. Explain the differences between standardized tests and informal objective examinations.
2. What reasons can you advance for the general conclusion that there is no conflict between standardized tests and informal objective tests?
3. Discuss the major advantages of the informal objective test over the traditional or essay test.
4. Discuss the limitations sometimes claimed for the informal objective test.
5. Distinguish among various types of instructional outcomes and illustrate each.



6. Briefly comment upon the selection of content and general construction of the teacher-made objective test.
7. Discuss pro and con the advisability of using several types of objective test items in the same classroom test.
8. Why should an objective test be so difficult that no pupil can make a perfect score and yet so easy that no pupil will have a zero score?
9. What cautions should be observed in administering the teacher-made objective test?
10. How should the various types of objective test items ordinarily be scored?
11. What procedures are useful to the teacher in the revision of the informal objective examination?
12. What are the major uses of the informal objective test?
13. Clearly distinguish between recall and recognition item forms.
14. Distinguish between the two ordinary forms of recall items and illustrate each type.
15. Give examples of several alternate-response item types.
16. Show the differences among the ordinary multiple-choice, the multiple-response, and the best-answer item forms. Illustrate.
17. Which type of matching exercise, the balanced or the unbalanced, is preferable? Why?
18. Give some of the most important general suggestions for the construction of objective test items.

## Selected References

- ADKINS, DOROTHY C., AND OTHERS. *Construction and Analysis of Achievement Tests*. Washington, D. C.: U. S. Government Printing Office, 1947.
- BROOM, M. E. *Educational Measurements in the Elementary School*. New York: McGraw-Hill Book Co., Inc., 1939. Chapters 5-6.
- BROWNELL, WILLIAM A., chairman. *The Measurement of Understanding*. Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1946.
- CONRAD, HERBERT S. "The Experimental Tryout of Test Materials." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 8.
- COOK, WALTER W. "Achievement Tests." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 1461-78.
- EBEL, ROBERT L. "Writing the Test Item." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 7.

- ENGELHART, MAX D. "Examinations." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 407-14.
- ENGELHART, MAX D. "How Teachers Can Improve Their Tests." *Educational and Psychological Measurement*, 4:109-24; Summer 1944.
- ENGELHART, MAX D. "Unique Types of Achievement Test Exercises." *Psychometrika*, 7:103-15; June 1942.
- GERBERICH, J. RAYMOND. *A Guide to Achievement Test Construction: Specimen Objective Item Types*. New York: Longmans, Green and Co., 1953.
- GERBERICH, J. RAYMOND. "A Technique for Measuring the Ability To Evaluate Objective Test Items." *Journal of Educational Research*, 27:46-50; September 1933.
- GREENE, EDWARD B. *Measurements of Human Behavior*. Revised edition. New York: Odyssey Press, 1952. Chapter 4.
- GREENE, HARRY A., AND CRAWFORD, JOHN R. *Work-Book in Educational Measurements and Evaluation*. New York: Longmans, Green and Co., 1945. Unit 2.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. p. 43-57.
- LEE, J. MURRAY. *A Guide to Measurement in Secondary Schools*. New York: D. Appleton-Century Co., Inc., 1936. Chapters 10-11.
- LINDQUIST, E. F. "The Construction of Tests." *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Co., 1936. Chapter 3.
- LINDQUIST, E. F. "Preliminary Considerations in Objective Test Construction." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 5.
- MICHEELS, WILLIAM J., AND KARNES, M. RAY. *Measuring Educational Achievement*. New York: McGraw-Hill Book Co., Inc., 1950. Chapters 5-10.
- MOSIER, CHARLES I., MYERS, M. CLAIRE, AND PRICE, HELEN G. "Suggestions for the Construction of Multiple-Choice Test Items." *Educational and Psychological Measurement*, 5:261-71; Autumn 1945.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. Chapter 3.
- ODELL, C. W. *How To Improve Classroom Testing*. Dubuque, Iowa: Wm. C. Brown Co., 1953.
- REMMERS, H. H., AND GAGE, N. L. *Educational Measurement and Evaluation*. New York: Harper and Brothers, 1943. Chapter 9.
- RINSLAND, HENRY D. *Constructing Tests and Grading in Elementary and High School Subjects*. New York: Prentice-Hall, Inc., 1937.



- ROSS, C. C. *Measurement in Today's Schools*. Second edition. New York: Prentice-Hall, Inc., 1947. Chapters 4-5.
- SPAULDING, GERALDINE. "Reproducing the Test." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 11.
- THUT, I. N., AND GERBERICH, J. RAYMOND. *Foundations of Method for Secondary Schools*. New York: McGraw-Hill Book Co., Inc., 1949. p. 171-86.
- TRAVERS, ROBERT M. W. *How To Make Achievement Tests*. New York: Odyssey Press, 1950. Chapters 2-6.
- TRAXLER, ARTHUR E. "Administering and Scoring the Objective Test." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 10.
- TYLER, RALPH W. "Identification and Definition of the Objectives To Be Measured." *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Co., 1936. Chapter 1.
- VAUGHN, K. W. "Planning the Objective Test." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 6.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapter 20.
- WEITZMAN, ELLIS, AND McNAMARA, WALTER J. *Constructing Classroom Examinations*. Chicago: Science Research Associates, 1949. Chapters 2-4.

## ***Constructing and Using Performance Tests***

THIS CHAPTER presents a brief treatment of the following points in the construction and use of performance tests:

- A. Growth of interest in performance testing of achievement.
- B. Measurable characteristics of performance.
- C. Object tests and their functions.
- D. Types of measures of performance.
- E. Methods of evaluating products.
- F. Constructing performance tests.
- G. Using performance test results.

Teachers and other users of educational achievement tests have long been aware that results from paper-and-pencil tests of facts and information in an instructional field reveal only a part of the story of educational accomplishment. Admittedly such tests are easily given, are quickly scored, and are extremely useful in the classroom, but they are also limited to the degree that in many instructional areas results from conventional tests emphasizing facts and principles are not highly correlated with actual performance. The recognition of this fact points up one of the serious limitations in the validation of many otherwise excellent achievement measures, and makes quite obvious the need for performance tests to supplement other measures of achievement.

In one form or another, performance tests are utilized in all three areas of measurement and evaluation treated in this volume—intelli-



gence, personality, and achievement. Such tests are used in measuring general intelligence of illiterate persons and individuals having language handicaps of other types. They are also used in the measurement of aptitudes for various types of manual skills. Conduct or performance tests are also extensively used in the evaluation of a wide variety of personality characteristics and traits. This chapter is primarily concerned with the measurement of those physical and motor reactions that represent important behavioral and skill outcomes of learning and that are, therefore, evidences of educational achievement.

## 1 NATURE OF PERFORMANCE TESTS

### Development of performance testing

Although performance tests were used by primitive peoples in their tests and ceremonies preceding the induction of youth into adult society and by the Greeks and Spartans in their athletic games, the objective written test came into use many years before objective performance tests received much attention. Despite the fact that several of the earliest instruments for the objective measurement of educational achievement were quality and so-called product, or source, scales for handwriting, drawing, English composition, and spelling, the paper-and-pencil test largely dominated in objective measurement until some twenty years ago. With an increasing realization on the part of psychologists and educators that knowledge of facts and principles is not necessarily accompanied by skills in their use and application, attention was directed, or perhaps redirected, to objective procedures for the measurement of skills in functional situations.

Broadly speaking, every test is a performance test, whether the performance consists of oral responses to questions, written responses on an essay or an objective test, or the application of physical or motor skills in a certain test situation. However, paper-and-pencil tests of factual knowledge are not generally regarded as performance tests. Neither are oral and essay tests, for that matter, but to the degree that the pupil's skill in expression is evaluated they actually are performance tests even though manipulation as such is not involved. For the purposes of this chapter, however, performance tests

will be considered primarily as those requiring the use and often the manipulation of physical objects and the application of physical and motor skills in situations not restricted to oral and written responses.

### Measurable characteristics of performance

A distinction of importance is found among the measurable characteristics in the field of performance testing. These instructional outcomes appear to be largely in the areas of knowledges, concepts, understandings, skills, and applications.

The knowledges, concepts, and understandings serving as necessary background for many types of skill performances are measured by what are variously called object tests, recognition tests, and identification tests. Recognition tests and identification tests do not necessarily imply the use of physical objects in the test situation but instead may involve photographic or drawn representations of the articles. These testing techniques most appropriately are considered in the treatment of objective written tests. Object tests, on the other hand, imply the presentation and use in the test situation of three-dimensional articles. Accordingly, they are dealt with in this chapter.

The skills and applications outcomes are measurable in some instances by written tests and in others by performance measures. Such skills and applications as those involved in reading comprehension, written expression, and arithmetic and mathematics are commonly evaluated by means of paper-and-pencil tests. So are some of the mathematical aspects of the sciences and even of the social studies. The aspects of performance testing for the direct measurement of skills and applications to be dealt with in this chapter are concerned with the procedures followed in performing a certain task and the product resulting from the completion of the task. Check lists and timing devices are the most widely used educational tools for evaluating the performance of the pupil, whereas quality scales, rating scales, score cards, and counting and measuring are commonly used in the evaluation of performance as it is evidenced in the completed product.

Tests of performance may be classified in several useful ways. The one chosen here divides the instruments and techniques for performance testing of educational achievement into: (1) object tests, (2) performance measures, and (3) product evaluations.



## 2 OBJECT TESTS

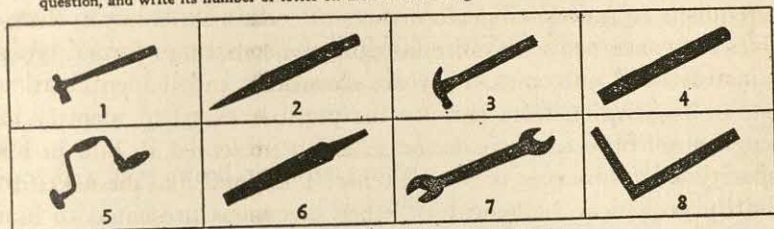
Object tests measure knowledges, concepts, and understandings prerequisite to functional performances of certain skills but in themselves are concerned with quite tangible and sometimes formal types of instructional outcomes. They are sometimes called identification tests or recognition tests because the pupil is asked to identify or recognize an object, specimen, or selection presented to him in his capacity as an observer or as a listener. The pupil may be asked to identify geological, biological, or other specimens presented to him in actuality or in the form of photographs or line drawings. He may be asked to recognize musical selections played by a soloist or orchestra or reproduced by a phonograph.

Although the visual and auditory senses are the ones primarily involved in such situations, the sense of touch may also enter in the case of objects where knowledge concerning grain, texture, or other surface qualities might aid in identification. The other two of the five basic senses—taste and smell—may even be employed in some less usual situations in the practical arts and physical sciences where physical objects are to be identified. The object test is more functional than the comparable type of test in which only photographic or drawn representations of the objects are presented, for the object may variously be seen, felt, listened to, and even smelled or tasted, whereas a pictorial representation can be interpreted only in terms of its visual stimulus.

The accompanying illustration from the *Prognostic Test of Mechanical Abilities* is used to represent an object test, although it actually involves the photographic presentation of objects and written responses by the pupils. However, if the student visualizes a setting in which the eight tools, appropriately numbered, are actually laid out on a table or bench and if he changes the first questions of the illustration so that they relate to the actual objects presented, he will obtain a clear idea concerning the setting and nature of an object test. The first three questions of the illustration are of the identification or recognition type and measure factual knowledges, but the last three questions go beyond formal knowledges in measuring concepts and understandings concerning the nature and appropriate uses of the tools.

Excerpts from Prognostic Test of Mechanical Abilities <sup>1</sup>

31-50. DIRECTIONS: Each of the following incomplete statements or questions is followed by five possible answers. For each item, select the answer that best completes the statement or answers the question, and write its number or letter on the line to the right.



Statements 31-45 refer to pictures of tools above

- |  |   |   |   |   |   |       |
|--|---|---|---|---|---|-------|
| 31. A claw hammer is shown in picture      | 1   | 3 | 6 | 7 | 8 | ___31 |
| 32. A chisel is shown in picture           | 2   | 4 | 5 | 6 | 7 | ___32 |
| 33. A ball peen hammer is shown in picture | 1   | 3 | 5 | 6 | 8 | ___33 |
| 39. Tool No. 1 can be used to:             | <input type="checkbox"/> file metal <input type="checkbox"/> polish metal <input type="checkbox"/> drill holes<br><input type="checkbox"/> take dents out of metal <input type="checkbox"/> caulk metal |   |   |   |   | ___39 |
| 40. Tool No. 2 can be used to:             | <input type="checkbox"/> mark metal <input type="checkbox"/> drive a screw <input type="checkbox"/> file metal<br><input type="checkbox"/> fasten a bolt <input type="checkbox"/> lock a nut            |   |   |   |   | ___40 |
| 41. Tool No. 3 can be used to:             | <input type="checkbox"/> cut wood <input type="checkbox"/> pull out a nail <input type="checkbox"/> bend a rod<br><input type="checkbox"/> drive a nut <input type="checkbox"/> tighten a nut           |   |   |   |   | ___41 |

Another illustration of an object test is drawn from the field of home economics. The accompanying illustration shows how pupils used actual shirts, ties, handkerchiefs, and socks displayed on a screen in demonstrating certain concepts and understandings concerning the significance of color in clothing selection.

Sample of Test in Clothing <sup>2</sup>

Assume that the articles displayed on Screen I are to be worn with a suit of dark-value gray, a top coat of middle-value gray, and middle-value pigskin gloves. Choose the most becoming shirt, tie, handkerchief, and socks for each man to wear with the gray suit, coat, and gloves. Write the number corresponding to your choice in the blank at the left of each item, and list no article more than once.

<sup>1</sup> J. Wayne Wrightstone and Charles E. O'Toole, *Prognostic Test of Mechanical Abilities*, Form A. Published by California Test Bureau, 1946.

<sup>2</sup> Clara B. Army, *Evaluation in Home Economics*. Appleton-Century-Crofts, Inc., New York, 1953. p. 143.



<i>Articles of Clothing</i>	<i>Descriptions of Men</i>
_____ 1. shirt	A. black hair, fair skin, and blue eyes
_____ 2. tie	
_____ 3. handkerchief	
_____ 4. socks	
_____ 5. shirt	B. medium brown hair, blue eyes, and somewhat sallow skin
_____ 6. tie	
_____ 7. handkerchief	
_____ 8. socks	
_____ 9. shirt	C. auburn (red) hair, brown eyes, and florid complexion
_____ 10. tie	
_____ 11. handkerchief	
_____ 12. socks	

### 3 PERFORMANCE MEASURES

The process by which a pupil produces some type of desired result in a test situation is appropriately observed and evaluated as to quality by use of check lists and as to quantity or time by use of a timing device. Performance measurement is often highly diagnostic in its significance, but it is time-consuming in those instances in which each pupil must be tested individually.

#### Check lists

The processes involved in the performance of a complex skill are subject to measurement and evaluation by the use of observational techniques. Although observation unaided by any objective instrument may often be effective when the observer is a qualified specialist in the skill in question, the use of check lists insures a more accurate and comprehensive record of the actual behavior of the individual observed. Such check lists have been evolved for a number of skill performances in industrial arts, home economics, and laboratory science.

Tyler illustrated procedures measurement in testing ability to use the microscope. The technique is necessarily used with only one person at a time because it requires full-time observation of the pupils by the examiner. The check list illustrated herewith includes a sequential listing of the appropriate and inappropriate steps of procedure in adjusting the instrument and finding a yeast cell or a blood cell on a slide, using a culture previously prepared by the examiner.

As the student prepares the slide, endeavors to adjust the microscope, and attempts to locate a cell, the examiner records his operations by numbers in sequence at appropriate places on the check list. The result is a diagnostic version of the student's success or failure in the assigned task. The numbers in the illustration show the results of such a record for an actual performance by a student.

STUDENT'S ACTIONS	Sequence of Actions	STUDENT'S ACTIONS (Continued)	Sequence of Actions
a. Takes slide	<u>1</u>	ag. With eye away from eyepiece turns down fine adjustment a great distance	<u>15</u>
b. Wipes slide with lens paper	<u>2</u>	ah. Turns up fine adjustment screw a great distance	
c. Wipes slide with cloth		ai. Turns fine adjustment screw a few turns	<u>16</u>
d. Wipes slide with finger		aj. Removes slide from stage	
e. Moves bottle of culture along the table		ak. Wipes objective with lens paper	
f. Places drop or two of culture on slide	<u>3</u>	al. Wipes objective with cloth	
g. Adds more culture		am. Wipes objective with finger	<u>17</u>
h. Adds few drops of water	<u>4</u>	an. Wipes eyepiece with lens paper	
i. Hunts for cover glasses	<u>5</u>	ao. Wipes eyepiece with cloth	
j. Wipes cover glass with lens paper		ap. Wipes eyepiece with finger	<u>18</u>
k. Wipes cover with cloth		aq. Makes another mount	
l. Wipes cover with finger		ar. Takes another microscope	
m. Adjusts cover with finger		as. Finds object	
n. Wipes off surplus fluid	<u>6</u>	at. Pauses for an interval	
o. Places slide on stage		au. Asks, "What do you want me to do?"	
p. Looks through eyepiece with right eye	<u>7</u>	av. Asks whether to use high power	
q. Looks through eyepiece with left eye	<u>9</u>	aw. Says, "I'm satisfied"	
r. Turns to objective of lowest power	<u>21</u>	ax. Says that the mount is all right for his eye	
s. Turns to low-power objective	<u>8</u>	ay. Says he cannot do it	<u>19, 24</u>
t. Turns to high-power objective		az. Told to start new mount	
u. Holds one eye closed		aa. Directed to find object under low power	<u>20</u>
v. Looks for light		ab. Directed to find object under high power	
w. Adjusts concave mirror		NOTICEABLE CHARACTERISTICS OF STUDENT'S BEHAVIOR	
x. Adjusts plane mirror		a. Awkward in movements	
y. Adjusts diaphragm	<u>10</u>	b. Obviously dexterous in movements	
z. Does not touch diaphragm		c. Slow and deliberate	<input checked="" type="checkbox"/>
aa. With eye at eyepiece turns down coarse adjustment	<u>11</u>	d. Very rapid	
ab. Breaks cover glass	<u>12</u>	e. Fingers tremble	
ac. Breaks slide		f. Obviously perturbed	
ad. With eye away from eyepiece turns down coarse adjustment		g. Obviously angry	
ae. Turns up coarse adjustment a great distance	<u>13, 22</u>	h. Does not take work seriously	
af. With eye at eyepiece turns down fine adjustment a great distance	<u>14, 23</u>	i. Unable to work without specific directions	<input checked="" type="checkbox"/>
		j. Obviously satisfied with his unsuccessful efforts	<input checked="" type="checkbox"/>
SKILLS IN WHICH STUDENT NEEDS FURTHER TRAINING	Sequence of Actions	CHARACTERIZATION OF THE STUDENT'S MOUNT	Sequence of Actions
a. In cleaning objective	<input checked="" type="checkbox"/>	a. Poor light	<input checked="" type="checkbox"/>
b. In cleaning eyepiece	<input checked="" type="checkbox"/>	b. Poor focus	
c. In focusing low power	<input checked="" type="checkbox"/>	c. Excellent mount	
d. In focusing high power	<input checked="" type="checkbox"/>	d. Good mount	
e. In adjusting mirror	<input checked="" type="checkbox"/>	e. Fair mount	
f. In using diaphragm	<input checked="" type="checkbox"/>	f. Poor mount	
g. In keeping both eyes open	<input checked="" type="checkbox"/>	g. Very poor mount	
h. In protecting slide and objective from breaking by careless focusing	<u>1</u>	h. Nothing in view but a thread in his eyepiece	
		i. Something on objective	
		j. Smears lens	<input checked="" type="checkbox"/>
		k. Unable to find object	<input checked="" type="checkbox"/>

Fig. 11. Check list of student reactions in finding an object under a microscope<sup>3</sup>

<sup>3</sup> Ralph W. Tyler, "A Test of Skill in Using a Microscope." *Educational Research Bulletin*, 9:493-96; November 19, 1930.



Supplementary portions of the check list provide for checking characteristics of student behavior, characteristics of the mount, and skills in which the student needs further training. When the student's performance is summarized and diagnosed by the use of these sections of the check list, the special type of remediation needed is ordinarily disclosed and serves as the basis for providing the necessary type of remedial instruction. The diagnostic significance of this type of summary is shown by the check marks in the illustration representing characteristics of and deficiencies in the student's performance.

As this individual technique is time-consuming, students can first be tested in a group situation where success or failure in a task common to all can be readily checked for each student by the quality of the adjustment he obtains. It then becomes necessary to use the check list only with those individuals who do not succeed in the group test situation.

This illustration is representative of the work-sample tests employed both in procedures and product measurement. The use of a microscope in finding a yeast cell or a blood cell may be considered as a sample, or work-sample, of the various ways in which a microscope is used in biological science.

## Timing devices

A stop watch or even an ordinary watch having a second hand is the only timing device the teacher ordinarily needs in performance testing, although some specialized types of performance may require the use of more precise timing instruments. In performances where speed is an important characteristic, the time required for the performance of the assigned task may constitute one measure, although in some instances there may be other and even more important measures. In the measurement of speed and accuracy in clerical, typing, stenographic, and other types of performances where speed is an important factor, a watch becomes a tool for the measurement of educational achievement.

It was pointed out in Chapter 3 that speed may be measured in terms of the amount or quantity of production in a given period of time or in terms of the time required to complete a product of a certain quality or to perform a task of a certain level of difficulty. This second method is the one ordinarily applied in performance testing when the job is to produce a completed product, such as a

planed and squared up board in manual arts or a seam in home economics. In a skill such as typing, however, the time is usually held constant and the quantity, as well as the quality, of the production is measured.

#### 4 PRODUCT EVALUATION

The quality of a skill performance can be evaluated in terms of the characteristics of the completed product as well as by the characteristics of the techniques used in its production. In assessing the characteristics of the product, quality scales, rating scales, score cards, and counting and measuring techniques are usually employed. There are occasions when it may be desirable to evaluate both the procedure and the product, but in many instances a good product depends so much upon effective procedures that product measurement alone may suffice. When such is the case, observation of each pupil separately during the performance ceases to be necessary. The resulting products can later be evaluated individually by the teacher.

##### Quality scales

Although standardized quality scales have been devised and used at least to some extent for measuring a variety of skills in composition, fine arts, industrial arts, home economics, and handwriting, it is probably in the last-mentioned area that they have been used most. The handwriting scale of the *California Achievement Test* is shown in Figure 12 to illustrate the quality scale. The pupils taking this handwriting test write the words used as samples on the scale. Each pupil's handwriting quality is then evaluated by finding the scale sample most closely resembling his writing and assigning the appropriate grade or age equivalent.

##### Rating scales and score cards

Rating scales and score cards are very similar devices for use in measuring the quality of the product made in a test situation. A numerical scale typically provides for separate ratings or evaluations on the various elements of skill required in the total performance. Such distinctive features of the total performance may range from only a few to a large number, depending on the complexity of the skill performance and the degree of analysis desired



in the evaluation. Numerical values representing various levels of quality usually range from three or four to ten. It is doubtful if more than ten degrees of quality can be distinguished reliably in evaluating qualitative performances.

grocery doubt concert	G. P. Age Equiv. (in months)	motion arrive believe
grocery doubt concert	3.0 99	
grocery doubt concert	3.5 105	motion arrive believe
grocery doubt concert	4.0 111	
grocery doubt concert	5.0 123	motion arrive believe
grocery doubt concert	6.0 136	
grocery doubt concert	7.0 148	motion arrive believe
grocery doubt concert	8.0 160	
grocery doubt concert	9.0 172	motion arrive believe
grocery doubt concert	10.0 184	
grocery doubt concert	10.5 190	motion arrive believe
grocery doubt concert	11.0 196	
	11.5+ 201	

Fig. 12. Handwriting scale of the California Achievement Tests <sup>4</sup>

<sup>4</sup> Ernest W. Tiegs and Willis W. Clark, *California Language Test Manual*, Intermediate. California Test Bureau, Los Angeles, 1950. p. 15.

Two samples are presented here to illustrate this method of product measurement. The first provides for a rather highly analytic

### Excerpt from Rating Form for Fastening <sup>5</sup>

(a) Nails:

(1) Straightness	1	2	3	4	5	6	7	8	9	10
	Are nails driven straight, heads square with wood, no evidence of bending?									
(2) Hammer marks	1	2	3	4	5	6	7	8	9	10
	Is wood free of hammer marks around nails?									
(3) Splitting	1	2	3	4	5	6	7	8	9	10
	Is wood free of splits radiating from nail holes?									
(4) Depth	1	2	3	4	5	6	7	8	9	10
	Are depths of nails uniform and of pleasing appearance?									
(5) Spacing	1	2	3	4	5	6	7	8	9	10
	Are nails spaced too close or too far apart?									
(6) Utility	1	2	3	4	5	6	7	8	9	10
	Will the nails hold?									

rating of a fairly simple skill. The second shows how a more complex skill may be evaluated in a less highly analytic manner.

### Food Score Card for Waffles <sup>6</sup>

WAFFLES				53
	1	2	3	Score
Appearance	1. Irregular shape	Regular shape		1.
Color	2. Dark brown or pale	Uniform, golden brown		2.
Moisture Content	3. Soggy interior or too dry	Slightly moist interior		3.
Lightness	4. Heavy	Light		4.
Tenderness	5. Tough or hard	Tender; crisp crust		5.
Taste and Flavor	6. Too sweet or flat or taste of leavening agent or fat	Pleasing flavor		6.
				SCORE

<sup>5</sup> Dorothy C. Adkins, *Construction and Analysis of Achievement Tests*. U. S. Government Printing Office, Washington, D. C., 1947. p. 231.

<sup>6</sup> Clara M. Brown, *Food Score Cards: Waffles*, No. 53. Published by University of Minnesota Press, 1940.



## Counting and measuring techniques

Counting most often becomes a direct product measurement technique when the quantity of articles produced in a given time or the errors made in a certain piece of work are important to the total evaluation. Speed of production is usually not greatly stressed in the school classroom, although where boys and girls are receiving training for certain types of employment, as they often are in trade and industrial schools, speed of performance may assume considerable significance. The error count in typewriting is a direct measure of quality in the product, and it is often combined as a penalty with the number of words typed per minute to obtain an evaluation of the total product. This procedure in effect provides a combined quantitative and qualitative score.

Measuring instruments of various types are also used in evaluating the quality of a product. Such devices as rules, calipers, squares, scales, gauges, and other instruments may be used in determining how accurately the pupil has performed the assigned task. Special mechanical testing devices may even be devised by the teacher of a skills subject to serve certain specific purposes.

Newkirk and Greene illustrated a performance test in which the product is measured objectively to determine the quality of pupil workmanship on a test of accuracy in woodworking.<sup>7</sup> After each pupil has been assigned to a work bench, the examiner reads the following instructions aloud.

*Directions to Pupil.* This is a test to determine how accurately you can use woodworking tools. The wood and all necessary tools will be given to you. The surfaces of the block of wood are numbered 1, 2, 3, 4, 5, 6. You will be given specific directions for doing the job and a working drawing that gives all the necessary dimensions. Do this project as accurately as you can. Do not waste time, but do not work too fast to do your best work. The steps must be done in the order given. After you begin work do not ask unnecessary questions, but if you are in doubt about a step in the procedure or a dimension on the working drawing ask the examiner. Write your *name* and *grade* in school on surface *No. 6* of the test block. *Do not begin work until the examiner gives the signal.*

<sup>7</sup> Louis V. Newkirk and Harry A. Greene, *Tests and Measurements in Industrial Education*. John Wiley and Sons, Inc., New York, 1935. p. 147.

*Procedure:*

1. Select face No. 1. Plane it square and true and to the thickness indicated on the working drawing. When finished re-mark No. 1.
2. Select side No. 2. Plane it square and true to surface No. 1. When finished re-mark No. 2.
3. Select end No. 3. Plane it square and true to No. 1 and 2. When finished re-mark No. 3.
4. Measure from end No. 3 toward end No. 5, and square a sharp pencil line across the block to the length indicated in the working drawing. Saw off the waste material with a back saw so that the stock will be as nearly the required length as you can make it. *Do not plane*. Re-mark end No. 5.
5. From edge No. 2 gauge a line the length of the block, allowing the exact width as indicated on the working drawing. Rip as nearly the exact width as possible. *Do not plane*.
6. On surface No. 1 lay out the center for the hole and bore.
7. When you have finished take your block to the examiner.

The examiner then supplies a copy of the working drawing, reproduced in Figure 13, to each pupil and makes sure that he has the

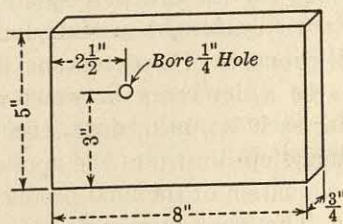


Fig. 13. Working drawing of a wood block <sup>8</sup>

necessary tools and a piece of standard wood stock prepared in advance by the instructor. The directions to the examiner, given last here for purposes of simplicity of presentation but obviously familiar to the examiner in advance, serve to illustrate the remaining steps in test administration.<sup>9</sup>

*Directions to Examiner:* It is essential that the pupil shall understand the exact procedure, and that he be able to visualize how the block is to look when finished. The following directions are recommended:

<sup>8</sup> *Ibid.* p. 146.

<sup>9</sup> *Ibid.* p. 146-47.



1. Read aloud and distinctly the directions to the pupil while the class follows silently. Answer any questions about the directions at this point.

2. Show the pupils a completed test block, and if they care to, let them examine it.

3. When there are no further questions, say, "Get ready. Hold up the test block. Begin work."

4. During the examination answer any questions about the steps in the procedure by rereading the step in question with the pupil.

5. Observe the pupils as they work to make certain that they are doing all the steps in the correct order.

6. Make certain that the proper tool is used where indicated, but do not tell the pupil how to use the tool.

7. Help any pupil having difficulty in interpreting the working drawing, but do not make any measurements on the test block for the pupil. This test measures ability to measure to  $\frac{1}{16}$  in. with a ruler, but is not a measure of ability to read drawings.

8. Take in the test block when the pupil has finished. The time is not important, for this is a test of quality or accuracy as it applies to modifying wood with simple hand tools.

The authors specified the use of a try square, a  $\frac{1}{4}$ -inch dowel 3 inches long, and a scale graduated in sixty-fourths of an inch in evaluating the pupil products. They suggested that for each rated dimension 10 points be assigned for an exact measurement and 1 point be deducted for each  $\frac{1}{16}$  inch of deviation from the specified dimension.<sup>10</sup> This final step illustrates the application of measuring instruments to the evaluation of the final product.

Army illustrated a functional situation in which each home economics pupil is given a miniature dress pattern such as that shown in Figure 14, a piece of paper to represent cloth from which a dress might be made, and the other necessary tools and materials used in the actual process of cutting dress material from a pattern and preparing it for sewing. To simulate conditions in which a dress would actually be made, it was recommended that the paper used in lieu of dress fabric have a design on one side, to represent the right side of dress material, that its length be three times its width, to represent three yards of cloth thirty-six inches wide, and, of course, that it be appropriate in size to the miniature pattern. Each girl would be expected to <sup>11</sup>

<sup>10</sup> *Ibid.* p. 147-48.

<sup>11</sup> Clara B. Army, *Evaluation in Home Economics*. Appleton-Century-Crofts, Inc., New York, 1953. p. 84.

cut out the pieces of the pattern and pin them on the paper strip, and then to draw the outline of the other half of each piece of the pattern in its proper location.

The resulting products could be evaluated by the teacher, using a rating scale or score card prepared specifically for this type of skill performance. This illustration of product measurement is of the simulated conditions or miniature type. When it is not practicable to employ the real situation in which a certain functional skill is employed, it is sometimes possible, as here, to simulate the conditions by the use of a miniature test representing the real situation quite accurately.

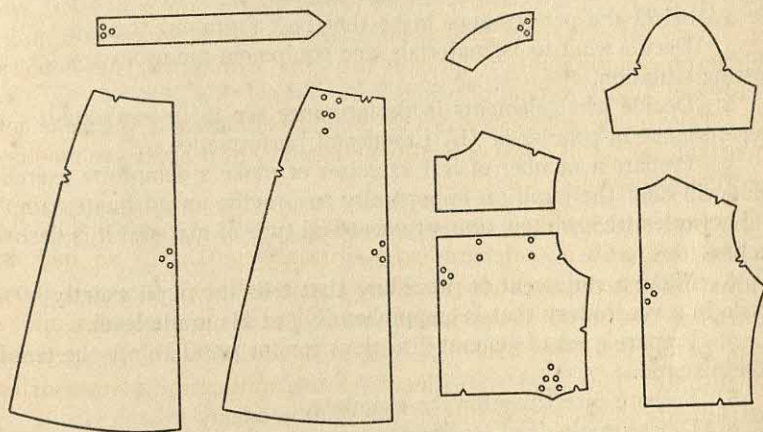


Fig. 14. Diagram of pieces of a dress pattern <sup>12</sup>

## 5 CONSTRUCTING PERFORMANCE TESTS

While performance and other types of manipulative tests have been widely used in certain educational fields, such as the industrial arts and home economics, the practical reliability of many of these devices has not been very satisfactory. It is believed that a part of this difficulty arises from the fact that too many of the better-known paper-and-pencil testing techniques have been uncritically borrowed and used without the necessary technical and administrative modifications required for effective testing in the specialized field. It is not

<sup>12</sup> *Ibid.* p. 85.



possible in this brief chapter to present examples of performance tests in many different areas of educational achievement, but a brief summary of certain general procedures that appear to be necessary in the construction of performance tests may be helpful.

The following summary of steps in preparing performance tests is taken with only minor modifications from a statement by Newkirk and Greene.<sup>13</sup> While these suggested steps were originally designed for use in classes in industrial education, there is little reason why they may not be expected to function equally well in other fields in which performance testing is desirable:

1. Make a job analysis of the activities covered in the course of study to determine exactly what qualities may be tested.
2. Select the performance tasks that best represent the job.
3. Decide what tools, materials, and equipment are necessary for the testing situation.
4. Decide what elements in performance are to be evaluated: (a) performance in process or (b) product of performance.
5. Prepare a number of test exercises or make a composite exercise that will offer the pupil an opportunity to provide an adequate sample of his work with each tool or instrument and type of material it is desired to test.
6. Make a statement of procedure that tells the pupil exactly what to do in a vocabulary that is comprehensible at his grade level.
7. Prepare a set of general directions for the pupil before the test is administered.
8. Prepare directions for the examiner.
9. Devise methods of scoring the test or evaluating the product that will provide an adequate measure of the results of each tool or instrument.
10. Try out the test on a few students, and make the more obvious changes and corrections in its content or directions.
11. Make two or more approximately equal forms of the test.
12. Try out the test, and compute the reliability coefficient, standard deviation, and standard error of a score, along with the grade and percentile or other types of norms.

A simple illustration of the manner in which certain of these steps are used in the construction of a performance test is given on pages 210 to 212 of this chapter.

<sup>13</sup> Newkirk and Greene, *op. cit.* p. 145.

## 6 USING RESULTS OF PERFORMANCE TESTING

The problems of using and interpreting performance test results are so nearly identical with those involved in the use of results from other types of tests that they need no special treatment here. In fact, performance test results are important to the teacher primarily because they supplement other important measures of ability or accomplishment.

### Topics for Discussion

1. List and discuss specific ways in which performance tests of accomplishment supplement information provided by other objective achievement tests.
2. Why should performance test results be particularly valuable in the testing of intelligence and personality qualities?
3. Discuss the truth of the statement that "every test is a performance test."
4. What are the chief limitations of performance testing procedures?
5. Discuss the three major types of performance testing techniques.
6. Show how quality scales, such as those for handwriting, drawing, lettering, sewing, soldering, splicing, and other industrial arts areas, are essential to the effective measurement of performance.
7. Discuss in some detail the adequacy with which the twelve steps in constructing a performance test presented on page 214 are applied in the case of the example given in this chapter.
8. Set out the specifications for the construction, use, and interpretation of a performance test in some other field than that illustrated in this chapter.

### Selected References

- ADKINS, DOROTHY C. *Construction and Analysis of Achievement Tests*. Washington, D. C.: U. S. Government Printing Office, 1947. Chapter 5.
- ARNY, CLARA B. *Evaluation in Home Economics*. New York: Appleton-Century-Crofts, Inc., 1953. Chapter 7.
- BLOOM, SAMUEL L. "Identification Tests." *Industrial Arts and Vocational Education*, 34:65-66; February 1945.
- GIBSON, JAMES J., editor. *Motion Picture Testing and Research*. Army Air Forces Aviation Psychology Program, Research Reports, No. 7.



- Washington, D. C.: U. S. Government Printing Office, 1947. Chapter 6.
- GREENE, EDWARD B. *Measurements of Human Behavior*. Revised edition. New York: Odyssey Press, 1952. Chapter 9.
- MICHEELS, WILLIAM J., AND KARNES, M. RAY. *Measuring Educational Achievement*. New York: McGraw-Hill Book Co., Inc., 1950. Chapters 11-14.
- NEWKIRK, LOUIS V., AND GREENE, HARRY A. *Tests and Measurements in Industrial Education*. New York: John Wiley and Sons, Inc., 1935. p. 116-23, 144-48.
- REMMERS, H. H., AND GAGE, N. L. *Educational Measurement and Evaluation*. New York: Harper and Brothers, 1943. Chapter 11.
- RYANS, DAVID G., AND FREDERICKSEN, NORMAN. "Performance Tests of Educational Achievement." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 12.
- SIRO, EINAR E. "Performance Tests and Objective Observation." *Industrial Arts and Vocational Education*, 32:162-65; April 1943.
- SUPER, DONALD E. *Appraising Vocational Fitness by Means of Psychological Tests*. New York: Harper and Brothers, 1949. p. 156-61.
- TYLER, RALPH W. "A Test of Skill in Using a Microscope." *Educational Research Bulletin*, 9:493-96; November 19, 1930.

## *Constructing and Using Evaluation Tools and Techniques*

THIS CHAPTER presents a treatment of the following aspects of evaluative tools and techniques:

- A. The nature and characteristics of evaluation.
- B. Tests of abilities to interpret.
- C. Tests of practices and activities.
- D. Pupil profile, progress, and class analysis charts.
- E. Cumulative records and report cards.
- F. The interview and the questionnaire.
- G. Evaluation in the classroom.

Evaluation is usually thought of as a broadly inclusive term. Accordingly, all types of tests, non-test tools, and techniques used in pupil appraisal may be considered evaluative. However, many of these instruments and techniques are treated elsewhere in this volume—achievement tests in Chapters 5 to 8, intelligence and aptitude tests in Chapter 10, and personality measures in Chapter 11. The line of demarcation between evaluative instruments and achievement tests on the one hand and evaluative techniques and personality measures on the other hand is not definite. The attempt is made in this chapter, therefore, to deal only with those tests, tools, and techniques not treated elsewhere in this volume but deemed most appropriate for use in evaluating and appraising pupil achievement.



## 1 MEANING OF EVALUATION

Evaluation is a concept still relatively new to education. The term has been used to include appraisal of the school program, curriculum, and instructional materials, appraisal of the teacher, and appraisal of the school child. Its methods run the gamut from observation and testing to elaborate research techniques. Evaluation as dealt with here is concerned in the direct sense only with characterizations of the school child through testing, measuring, and appraising.

### Nature of evaluation

Wrightstone's definition exemplifies the point of view appropriate when the school child is the focus of evaluation.

Evaluation is a relatively new technical term, introduced to designate a more comprehensive concept of measurement than is implied in conventional tests and examinations . . . the emphasis in measurement is upon single aspects of subject-matter achievement or specific skills and abilities, but . . . the emphasis in evaluation is upon broad personality changes and major objectives of an educational program. These include not only subject-matter achievement but also attitudes, interests, ideals, ways of thinking, work habits, and personal and social adaptability.<sup>1</sup>

This definition supports the distinctions made in Chapter 1 among testing, measuring, and evaluating. The terms represent successively more inclusive and meaningful approaches to the appraisal of pupils. Evaluation thus includes not only the methods and tools appropriate in testing and measuring but also a variety of procedures and instruments of broader scope.

### Characteristics of evaluation

A somewhat more adequate understanding of evaluation can be obtained by considering its characteristics. Wrightstone characterized evaluation by stating its purposes and methods.

First, it attempts to measure a comprehensive range of objectives of the modern school curriculum rather than limited subject-matter achieve-

<sup>1</sup> J. Wayne Wrightstone, "Evaluation." *Encyclopedia of Educational Research*, Revised edition. Macmillan Co., New York, 1950. p. 403.

ment only. Second, modern evaluation uses a variety of techniques of appraisal such as achievement, attitude, personality and character tests . . . rating scales, questionnaires, judgment scales of products, interviews, controlled-observation techniques, sociometric techniques and anecdotal records. Third, evaluation includes integrating and interpreting the various indexes of behavior changes so as to construct an inclusive portrait of an individual. . . . For this purpose a comprehensive cumulative record is valuable.<sup>2</sup>

A somewhat more extensive characterization is that presented by Quillen and Hanna, who stated that:

1. Evaluation includes all the means of collecting evidence on student behavior.
2. Evaluation is more concerned with the growth which the student has made than with his status in the group. . . .
3. Evaluation is continuous . . . an integral part of all teaching and learning.
4. Evaluation is descriptive as well as quantitative.
5. Evaluation is concerned with the total personality of the student and with gathering evidence on all aspects of personality development.
6. Evaluation is a cooperative process involving students, teachers, and parents.<sup>3</sup>

These characterizations of evaluation justify the concern here with evaluation as distinguished from testing and measuring. For that reason the treatment of this chapter should be considered particularly in relation to Chapters 5 to 8.

## 2 EVALUATIVE TESTS

The distinction between a test that measures and a test that evaluates is by no means exact. It is too much to demand that a test meet all of the characteristics outlined in the above section of this chapter, for testing techniques have not yet provided, and in fact may never provide, single instruments so broadly conceived. The tests considered here to be evaluative may in general be distinguished from tests that measure by their greater emphasis on the less tangible

<sup>2</sup> J. Wayne Wrightstone, "Trends in Evaluation." *Educational Leadership*, 8:91-95; November 1950.

<sup>3</sup> James Quillen and Lavone A. Hanna, *Education for Social Competence*. Copyright by Scott, Foresman and Co., Chicago, 1948. p. 343-46. Reprinted by permission.



or even intangible types of instructional outcomes, on the types of outcomes resulting from broad and varied learning experiences, and on the ability of the pupil to apply and to use information in reasoning and problem-solving. Such tests are here considered as of two types: (1) interpretive tests and (2) tests of practices and activities. The tests illustrated and discussed below are only in small degree representative. The wide variety of techniques used and the length of adequate illustrations preclude a wider sample.

Most of the tests of these types available in published form are for use in the high school or college. This is perhaps because to some extent they embody the philosophy of general education, so far considered most applicable to levels above the intermediate grades. A few such instruments are available for use in the junior high school and intermediate grades, however. Moreover, many of these tests are not provided with norms, since norms for tests of quite intangible outcomes appear to lack precise meaning.

### Interpretive tests

Tests measuring abilities to interpret are similar to reading comprehension tests in that both include not only the test items but also the material to which the test items refer. This material is usually in verbal form for reading tests and often consists of a paragraph or short selection on which the test items are based. It may but often does not appear in verbal form for interpretive tests, as tabular and graphical materials are frequently the basis for the interpretations the pupils are asked to make. Furthermore, reading tests typically measure ability to answer questions of fact or to distinguish major ideas in the selection, whereas interpretive tests measure such complex abilities as are involved in the interpretation of data, application of scientific principles, logical reasoning, and critical thinking.

The accompanying excerpt from the lower level of the *Interpretation of Data Test*, designed for use in the junior and the senior high school, represents a rather complex test unit based on a chart. Pupils are instructed to respond by answering "T" if enough information is given to make the statement true, "F" if enough information is given to make the statement false, or "U," meaning uncertain, if insufficient information is given to warrant a decision. Items in the upper level of this test, designed for use in the senior high school and





Excerpts from Watson-Glaser Tests of Critical Thinking <sup>5</sup>

## VI. Can rich and poor obtain, on the whole, equal justice from the courts in the United States today?

21. No; for all governmental agencies in a capitalistic society are fundamentally designed to protect the privileges of the owning class.....	41	42
	Strong	Weak
22. No; there are many dramatic cases illustrating prejudice against the poor....	43	44
	Strong	Weak
23. Yes; judges take an oath to support the law and the Constitution without fear or special favors. ....	45	46
	Strong	Weak
24. Yes; when a poor man sues a rich man or a large corporation, the jury's sympathies are more likely to be with the poor man, thus balancing any other advantage which the rich man may have. ....	47	49
	Strong	Weak

17. John asserted that all races are alike in abilities; only differences in opportunity make some better educated, more artistic, more honest, or more successful than others. Jim answered, "John, you're crazy! The idea that Negroes, Indians, and Japanese all have the same ability and talent and character that white people do is ridiculous. Anyone with the least bit of common sense should not make such a foolish statement."

We may properly conclude that —

- 81 Jim believed that there are differences between the abilities of white people and the abilities of Negroes, Indians, and Japanese.
- 82 John failed to take account of the history of the accomplishments of the various races.
- 83 Jim understood the facts better than John did.
- 84 John understood the facts better than Jim did.
- 85 None of the above conclusions properly follows from the information given.

The last illustration of this type to be given here is from the *Logical Reasoning Test*, one of the evaluation instruments of the Eight-Year Study of the Progressive Education Association, designed for use in Grades 10 to 12. The directions to the pupils appearing in the two boxes of the accompanying excerpt indicate that the pupil is asked not only to choose the appropriate conclusion from the three provided but also to evaluate statements—four of the actual twelve appear in the excerpt—on their significance for the conclusion chosen.

<sup>5</sup> Goodwin Watson and Edward M. Glaser, *Watson-Glaser Tests of Critical Thinking*: (1) Test 3, Discrimination of Arguments, and (2) Test 8, Applied Logical Reasoning. Published by World Book Co., 1942.

Excerpt from Logical Reasoning Test<sup>6</sup>Problem IV

In 1940, W. Gibson Carey, Jr., then president of the Chamber of Commerce of the United States, made the following statements in an address delivered at Waukegan, Illinois:

"Our primary job is to kill excess government spending lest it kill us. Excess spending is inevitable when the central government takes over many of the duties which rightly belong to the states. The central government of the United States has taken over many of the duties which rightly belong to the states."

Directions: Examine the conclusions given below. If the statements made by Mr. Carey are true, which one of the conclusions do you think is justified?

Conclusions

- X. If the central government had not taken over many duties which rightly belong to the states, excess spending would have been avoided.
- Y. Since the central government took over many duties which rightly belong to the states, it was not possible to avoid excess spending.
- Z. Further information is needed before any logical conclusion can be drawn.

A: Statements which explain why your conclusion is logical.

Mark in column B: Statements which do not explain why your conclusion is logical.

C: Statements about which you are unable to decide.

Statements

- 1. Before any logical conclusion can be reached, one must know whether the state governments would be more efficient than the central, or federal, government.
- 2. If one removed the fundamental cause for excess spending, excess spending would be avoided.
- 3. Since we accept the statements made by Mr. Carey as true, excess spending could not be avoided if the central government took over many duties which rightly belong to the states.
- 4. If centralization of government always leads to excess spending, then when centralization of government occurs, excess spending will also occur.

## Tests of practices and activities

Paper-and-pencil tests of practices and activities must of necessity consist of verbalized statements or questions to which the pupils react instead of direct measurement. However, pupils' responses to tests of this type may well disclose information that can lead to inferences concerning their individual interests, personalities, and adjustment. In fact, such tests are similar in some respects to the adjustment inventories treated in Chapter 11.

A few items of the *Health Activities Inventory* are shown in an accompanying excerpt together with the instructions for Parts I and II. Part I is designed to obtain information concerning a student's participation in desirable and undesirable health activities, whereas Part II measures his estimate of the validity of certain health practices. This instrument is one of the six health inventories

<sup>6</sup> Eight-Year Study of the Progressive Education Association, *Logical Reasoning Test*. Published by Educational Testing Service, 1950.



developed by the Cooperative Study in General Education for use from Grades 9 to 16. Illustrations from the *Health Attitudes* and *Health Interests* inventories appear in Chapter 11 and a brief discussion of the entire series appears in Chapter 21.

### Excerpts from Health Activities Inventory <sup>7</sup>

#### Part I

**Directions:** The following list of 100 items represents some of the activities in which people engage or which may affect them. Read each item, and on the appropriate line of the answer sheet blacken the space under

- R if it is a practice which you follow regularly or if it happens to you frequently,  
 O if it is a practice which you engage in occasionally or if it happens to you occasionally,  
 N if it is a practice which you never engage in or if it never happens to you.

1. Squeeze a blister to remove its watery content.
2. Treat skin disorders (such as acne) with common drugstore preparations not prescribed by a physician.
3. Pick blackheads, pimples, etc., with a needle or some other sharp object.
4. Remove moles or warts yourself.
5. Use a salve or a liquid for bleaching the skin.

#### Part II

After you have answered these 100 items according to the directions for Part I, go back to the first item and, starting with space 101 on the answer sheet, mark the first 70 items again according to the following directions: Blacken the space under

- S if you believe the practice to be a sound one; that is, if you believe the practice is substantiated by science principles;  
 D if you are doubtful about the soundness of the practice; that is, if you are not sure whether the practice can be substantiated or is contradicted by science principles;  
 U if you believe the practice to be an unsound one; that is, contrary to science principles.

The *Inventory of Personal-Social Relationships* is illustrated herewith by the directions and a few sample items from Part I, on activities and interests. Part II deals in quite similar manner with the students' concerns and difficulties. This inventory, for use from Grade 9 through the college years, is designed to measure development of the individual student in the area of personal-social relationships. Brouwer outlined steps for summarizing group results and analyzing an individual student's responses on the parent instrument, only slightly revised in the present edition.<sup>8</sup>

<sup>7</sup> Cooperative Study in General Education, *Health Activities*, Health Inventory No. I. Published by Educational Testing Service, 1950.

<sup>8</sup> Paul J. Brouwer, *Student Personnel Services in General Education*. American Council on Education, Washington, D. C., 1949. p. 190-204.

Excerpt from Inventory of Personal-Social Relationships<sup>9</sup>

## PART I. ACTIVITIES AND INTERESTS

**Directions:**

Mark your responses as follows in the A U D column at the left-hand side of the page:

- (A) U D -- Draw a circle around the A if the item represents an activity in which you participate or something you do, either occasionally or frequently.
- A (U) D -- Draw a circle around the U if the item represents something you rarely or never do BUT in which you are interested--that is, something you would like to do.
- A U (D) -- Draw a circle around the D if the item represents something you rarely or never do AND toward which you are more or less indifferent.

DO NOT OMIT ANY ITEMS. IF YOU ARE DOUBTFUL ABOUT YOUR RESPONSE, MAKE THE BEST QUICK DECISION YOU CAN. DO NOT PAUSE TOO LONG ON ANY ONE STATEMENT

s	A U D	1. Going to a student "hangout" with friends for a Coke, a snack, etc.
k,i	A U D	2. Singing in a glee club, chorus, quartet, or similar musical group.
s,k	A U D	3. Playing on an organized athletic team (varsity or intramural).
f	A U D	4. Attending student-faculty teas.
s,k	A U D	5. Going to dances.

## 3 OTHER EVALUATIVE TOOLS

Evaluative tests of the types represented in the preceding section are of more recent origin than are most of the evaluative tools to receive consideration here. The evaluative significance of such tools as the pupil profile and pupil progress charts, the cumulative record, the report card, and the class analysis chart lies much more in the broadened conceptions employed in their construction and use than in their uniqueness. Each of these tools is briefly discussed below.

## Pupil profile chart

Pupil profile charts are provided with many standardized tests of general achievement and many diagnostic tests in order to show differences in achievement levels graphically. The charts, frequently providing places for various part and total scores and their graphical

<sup>9</sup> Cooperative Study in General Education, *Inventory of Personal-Social Relationships*. Published by Educational Testing Service, 1950.



representation, often appear on the front covers of the test booklets or on pupil answer sheets. It is often recommended by test authors that the answer sheet or the cover of the booklet, which also carry such information as the pupil's name, the name and form of the test used, and the date on which it was given, be filed in the pupil's cumulative record folder or elsewhere for future reference and use.

The following illustration of a pupil profile chart is representative of the charts provided with most general achievement tests. The illustration shows the method of using a profile for results from a

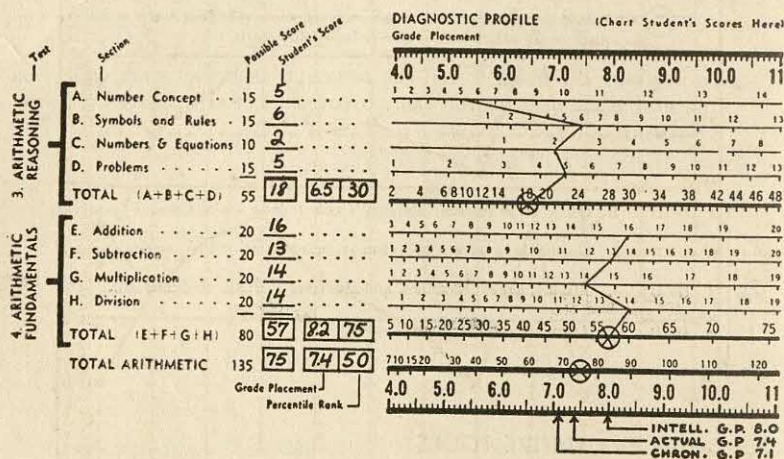


Fig. 15. Sample profile chart for the California Arithmetic Test<sup>10</sup>

single test that is part of a general achievement battery. It provides for the listing of scores on the achievement test and for a profile showing achievement levels on the total test, on its two major areas, and on its eight parts. Relative strengths and weaknesses appear graphically for ready observation by the teacher.

### Pupil progress chart

Evidence of pupil progress as measured by achievement tests over a period of years can be presented graphically by the use of a pupil profile chart. An illustration of this procedure is shown in Figure

<sup>10</sup> Ernest W. Tiegs and Willis W. Clark, *Manual for California Arithmetic Test, Intermediate*. California Test Bureau, Los Angeles, 1951. p. 4.

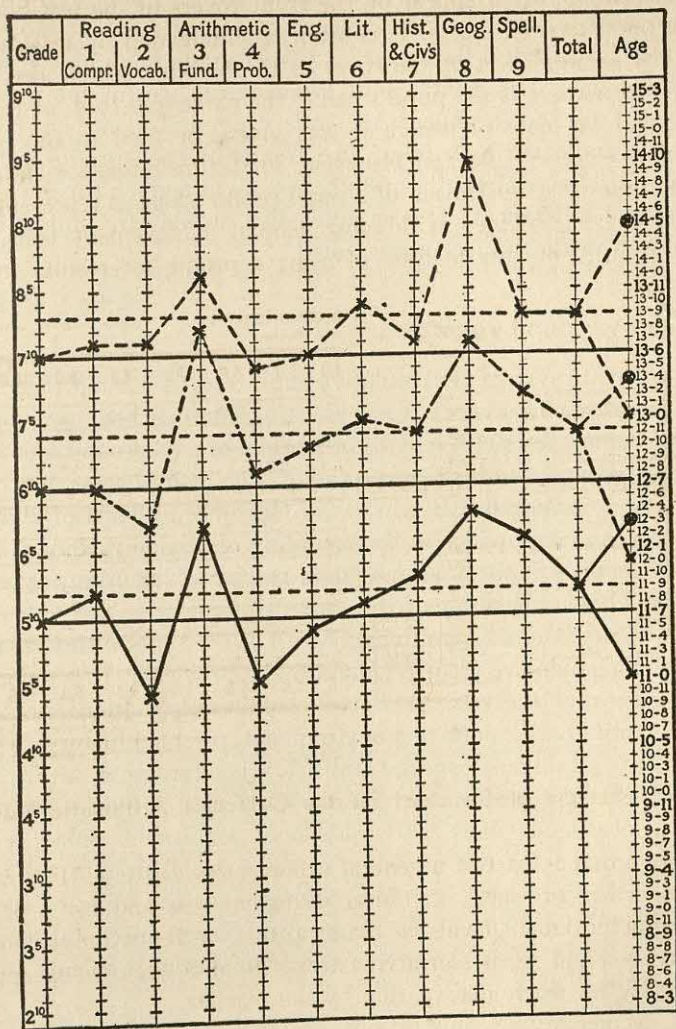


Fig. 16. Sample graphic record of pupil progress for the Metropolitan Achievement Tests <sup>11</sup>

<sup>11</sup> Richard D. Allen and others, *Supervisor's Manual: Metropolitan Achievement Tests*. World Book Co., Yonkers, N. Y., 1935. p. 32.



16 for a pupil tested for three successive years by the *Metropolitan Achievement Tests*. The pupil's average achievement is shown by the broken horizontal lines and his school grade level by the heavy horizontal lines. His mental and chronological ages at the times he was tested are shown under "Age," the years and months of mental age being indicated by circles. The trends of each profile indicate the pupil's relative strengths and weaknesses at a given time, whereas the vertical differences between successive profiles are indicative of his growth in the subject areas tested.

### Cumulative pupil record

An adequate system of cumulative pupil records is almost essential if the program of a school is to be effective. Many schools apparently keep no cumulative pupil records other than of background facts concerning the pupil and his parents and of scholastic success, but other schools have comprehensive and even elaborate systems of cumulative records that provide for the recording of a wide variety of data concerning each pupil in a cumulative record folder. Many types of variations between these extremes are also found.

No attempt is made here to catalog all of the types of information for which cumulative records should make provision. It is sufficient to indicate that the records should contain information about the pupil's family background and environment, personal history, health, personality, intelligence, special abilities, school progress, scholarship, achievement test performances, extra-curricular activities, employment, educational plans, and vocational ambitions. Some record systems provide for the recording of data on all or most of these points on a record card or folder, and also for the filing of certain types of other data, such as test profiles and scores, anecdotal records, case studies, and reports of action taken on special problems, in the cumulative record folder.

The accompanying illustrations of both sides of the *American Council on Education Cumulative Record for Elementary and Secondary Schools* show a form on which a pupil's record may be kept over a period of years. No attempt is made here to discuss the types of evidence for which provision is made, but an examination of the forms will reveal that a very comprehensive picture of a pupil can be obtained from the types of data that can be recorded on such a form.

It is impossible here to discuss at all adequately the values and uses of the cumulative record in pupil guidance and adjustment. However, it should be apparent that the mere availability of such a variety of information for all pupils in a school as can be recorded on the pictured type of cumulative record is of great value. Such records are useful to administrators, to guidance workers, and to teachers as a basis for careful analyses in cases of maladjustment or disciplinary difficulties, and on others of the many occasions requiring or at least making desirable comprehensive information about individual pupils.

### Pupil report card

Although the traditional report card may be considered to be an evaluative tool, the better modern report cards merit that designation much more definitely. The report card presents to the pupil and his parents a series of evaluations of his scholastic success and frequently of other aspects of his school performance. Because report cards are so widely known and because their organization and content differ so greatly, it seems neither desirable nor feasible to discuss further or to illustrate this evaluative tool here.

### Class analysis chart

Class analysis charts are valuable tools in the summarization of results from testing. Although such charts as are provided with standardized tests vary greatly, they usually provide a means of showing median achievement for the class or pupil group and the position of each pupil in the group in relation to age norms, grade norms, or both, for elementary-school tests. High-school tests more frequently provide for the graphical representation of median group performance and individual pupil status in relation to grade norms or percentile norms. The following illustration and discussion are based on a class analysis chart which is rather typical of those usually provided with general achievement test batteries.

The chart reproduced on pages 232 and 233 gives an analysis of the results from the use of the *Metropolitan Achievement Tests* with a class of 22 pupils in the second month of the sixth grade. Median class standing is shown by the crosses for achievement on the entire test and in the various subjects for which the test provides, as well as for intelligence quotients and chronological and mental ages. The line

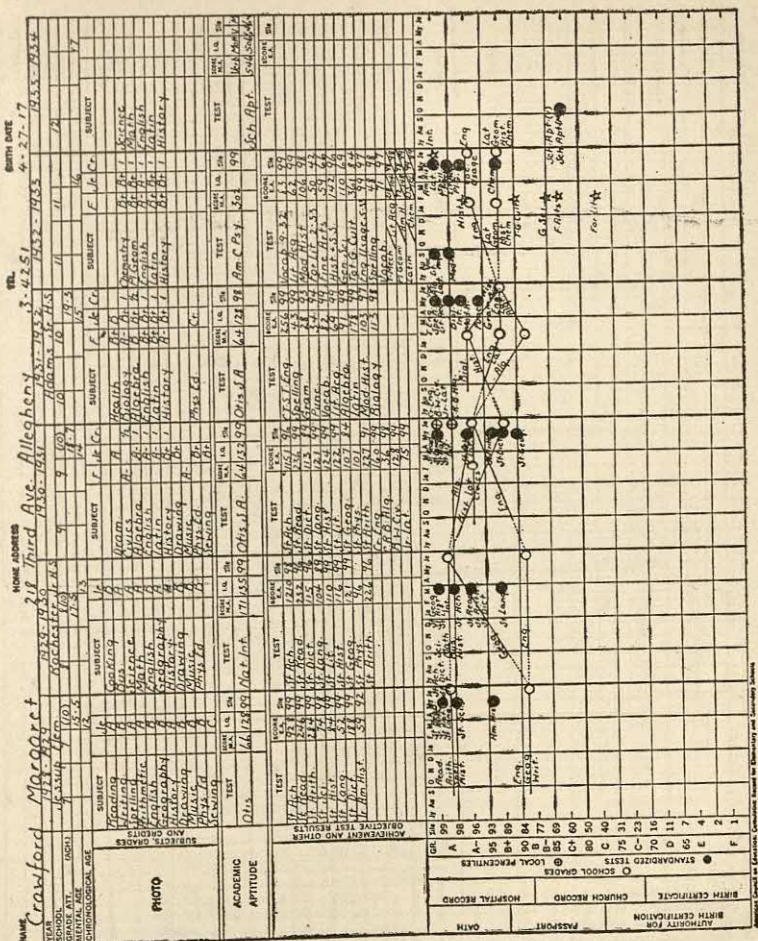


NAME GIVEN MIDDLE SURNAME	BIRTH DATE	PLACE BORN	GENERAL EDUCATION	RELIGION	RACE & NATIONALITY	ARRIVED IN U.S. DATE	OTHER	EDUCATION AMOUNT AND KIND	OCCUPATION
Winfred Margaret	1921-11-11	Allegheny						High School	Housekeeper
Alfred		Penna.	Good					Film	
Anna		Penna.	Good						
STEPHENSON									
LANGUAGE SPOKEN IN HOME									
AFTER 1910 English									
TYPE OF HOME COMMUNITY									
BEFORE 1910 City									
AFTER 1910 City									
YEAR AND AGE	1928-1929	1930-1931	1932-1933	1934-1935	1936-1937	1938-1939	1940-1941	1942-1943	1944-1945
BOYS	Donald	William	William	William	William	William	William	William	William
ATTENDANCE	AB.	AB.	AB.	AB.	AB.	AB.	AB.	AB.	AB.
DISCIPLINE	Good	Good	Good	Good	Good	Good	Good	Good	Good
INTELLECTUAL INFLUENCES AND SOCIAL ADJUSTMENT	Good	Good	Good	Good	Good	Good	Good	Good	Good
MENTAL AND EDUCATIONAL	Good	Good	Good	Good	Good	Good	Good	Good	Good
PHYSICAL AND ATHLETIC	Good	Good	Good	Good	Good	Good	Good	Good	Good
EXTRACURRICULAR AND FREE-TIME INTERESTS AND EXPERIENCES	Good 44-54 44-64 Basketball	Good 44-54 44-64 Basketball	Good 44-54 44-64 Basketball	Good 44-54 44-64 Basketball	Good 44-54 44-64 Basketball	Good 44-54 44-64 Basketball	Good 44-54 44-64 Basketball	Good 44-54 44-64 Basketball	Good 44-54 44-64 Basketball
EDUCATIONAL PLANS	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor
EDUCATIONAL PLANS	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor	Teacher Lawyer Doctor
PERSONALITY	1-2	1-2	1-2	1-2	1-2	1-2	1-2	1-2	1-2
RATINGS	0	0	0	0	0	0	0	0	0
REMARKS									

Fig. 17. Sample American Council on Education

connecting the crosses provides a profile of median achievement. Individual pupils are designated by identifying numbers placed at positions on the chart representing their scores. The distributions of intelligence quotients in column B and of mental ages in column E are related to but not really a part of the chart proper. At the bottom of the chart is shown the median achievement in terms of grade equivalents.

Among the significant interpretations possible from this type of chart are those involving comparisons of median grade equivalents





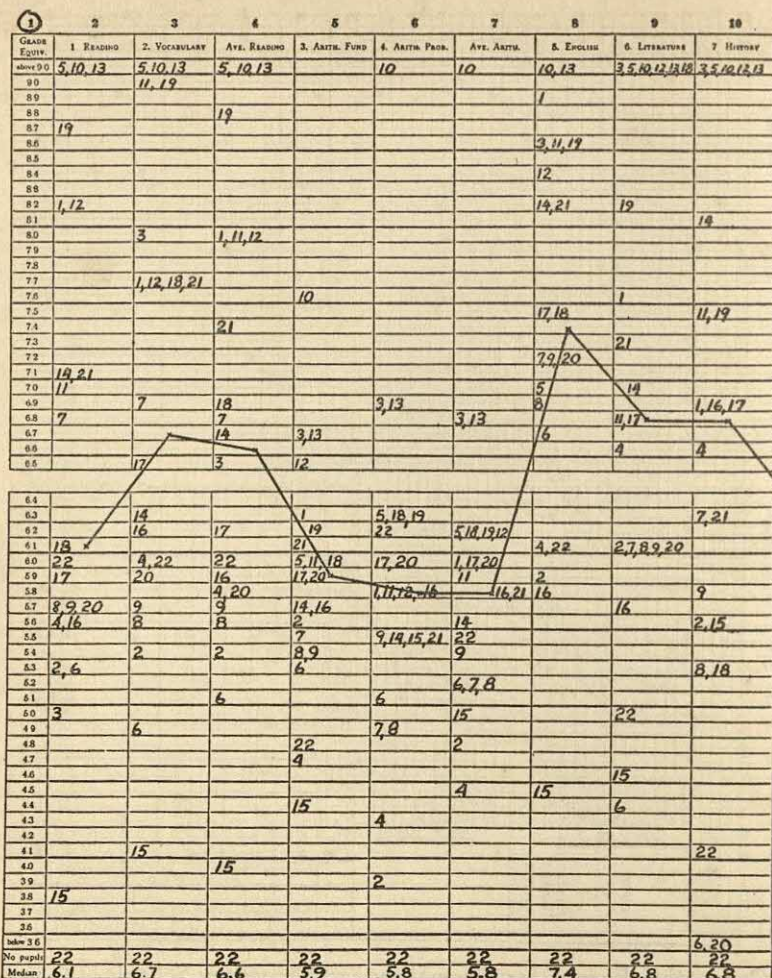


Fig. 18. Sample class analysis chart for

#### 4 EVALUATIVE TECHNIQUES

To be distinguished from tests and tools used in pupil evaluation are two major techniques used for the same general purpose: (1) the interview and (2) the questionnaire. The questionnaire is considered here, even though it is *per se* an instrument, because questionnaires typically are constructed to meet each need as it arises and are not found available in generalized published form.





child's interests, needs, and background about which the teacher needs information. Even in such informal uses of the interview as may be of concern to the teacher, it is essential for best results that rapport be established between the teacher and the child. A frightened or an antagonistic pupil is not a good subject for an interview. Therefore, the teacher should give the same type of attention to the establishment of rapport that is necessary prior to the administration of individual intelligence tests. Pupils should not be questioned on many types of issues in the presence of a third party, for their responses might then be less frank and spontaneous than if they were questioned in privacy.

In this broad sense, the interview is widely useful and flexible. However, it extends possibilities to the teacher for learning more about his pupils and consequently aids him in attempting to effect the best adjustment possible for each pupil.

### The questionnaire

Questionnaires have been very widely used, and all too frequently misused, in the attempt to evaluate some aspect of the school program and to measure certain intangible types of pupil behavior. Questions of fact appear in such instruments when they are used in obtaining simple, factual information or when they are used, for example, in obtaining information concerning pupil activities. Certain intangible tastes and preferences, primarily in the areas of attitudes and interests, are also measurable by the use of questionnaires. Adjustment inventories also typically make use of this technique. Part I of the *Health Activities Inventory*, treated in a preceding section of this chapter, and the several attitudes, interests, and adjustment inventories dealt with in Chapter 11 illustrate the use of this technique in the evaluation of pupil behavior.

## 5 EVALUATIVE TOOLS AND TECHNIQUES IN THE CLASSROOM

It should be apparent from the preceding discussion and illustrations that the equipment of the classroom teacher may well include evaluative tests, tools, and techniques as well as the more traditional types of measuring instruments and techniques. As has been made clear in Chapters 7 and 8, the classroom teacher very properly may construct informal objective tests and performance tests to meet his particular needs and to supplement standardized achievement tests

and scales. He can also participate in the cooperative planning and development of most of the evaluative instruments and techniques treated in this chapter. However, the development of some types of test units, such as those illustrated for the *Interpretation of Data Test* and the *Logical Reasoning Test*, involves principles of test construction going somewhat beyond those presented in Chapters 5 and 7 and entails a considerable amount of experience on the part of the test maker. As Ebel commented, "Skilled, experienced item writers find it difficult to construct interpretive exercises of high quality."<sup>14</sup>

The interpretation of evaluation results demands more insight and understanding than is ordinarily required for handling results from more traditional objective tests, and as yet few guides other than those provided with the specific instruments themselves have been set up as aids to users. Moreover, the broad, integrative nature of evaluation and the wide variety of instruments and techniques preclude definite limitations on the use of results. Therefore, the considered judgment of the evaluator not only of the direct results of evaluation but also of all other sources of information about individual pupils should be exercised in drawing conclusions and in deciding upon any indicated courses of action.

### Topics for Discussion

1. How are evaluative tests distinguishable from other paper-and-pencil achievement tests?
2. For what grade levels are evaluative tests most often provided at present?
3. How do interpretive tests differ from more traditional tests in purposes and testing techniques?
4. What are some major characteristics of tests of practices and activities?
5. How can a pupil profile chart serve as a pupil progress chart?
6. For what types of information should a cumulative pupil record make provision?
7. Describe a typical class analysis chart and discuss its uses by the teacher.
8. Briefly discuss the interview and the questionnaire as evaluative techniques.
9. How can the teacher make effective use of evaluative tests, tools, and techniques in the classroom?

<sup>14</sup> Robert L. Ebel, "Writing the Test Item." *Educational Measurement*. American Council on Education, Washington, D. C., 1951. p. 246.



## Selected References

- ALLEN, WENDELL C. *Cumulative Pupil Records*. New York: Bureau of Publications, Teachers College, Columbia University, 1943.
- BINGHAM, WALTER V., AND MOORE, BRUCE V. *How to Interview*. Revised edition. New York: Harper and Brothers, 1941.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 726-51.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p. 428-65.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938. p. 43-49.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. p. 644-83.
- DRISCOLL, GERTRUDE. *How To Study the Behavior of Children*. Practical Suggestions for Teaching, No. 2. New York: Bureau of Publications, Teachers College, Columbia University, 1941.
- ENGELHART, MAX D., AND LEWIS, HUGH B. "An Attempt To Measure Scientific Thinking." *Educational and Psychological Measurement*, 1:289-94; July 1941.
- FINDLEY, WARREN G. "Educational Evaluation—Recent Developments." *Social Education*, 14:206-10; May 1950.
- FINDLEY, WARREN G. "A Statistical Index of Participation in Discussion." *Journal of Educational Psychology*, 39:47-51; January 1948.
- FREEMAN, FRANK S. *Theory and Practice of Psychological Testing*. New York: Henry Holt and Co., 1950. Chapter 15.
- FROELICH, CLIFFORD P., AND DARLEY, JOHN G. *Studying Students: Guidance Methods of Individual Analysis*. Chicago: Science Research Associates, 1952. Chapters 4-8.
- GOODENOUGH, FLORENCE L. *Mental Testing: Its History, Principles, and Applications*. New York: Rinehart and Co., Inc., 1949. Chapter 26.
- HARTUNG, MAURICE L., AND OTHERS. "Aspects of Thinking." *Appraising and Recording Student Progress*. New York: Harper and Brothers, 1942. Chapter 2.
- KOOS, L. V. *The Questionnaire in Education*. New York: Macmillan Co., 1928.
- OSS Assessment Staff. *Assessment of Men: Selection of Personnel for the Office of Strategic Services*. New York: Rinehart and Co., Inc., 1948.
- RATHS, LOUIS. "A Thinking Test." *Educational Research Bulletin*, 23:72-75; March 15, 1944.

- Research Division, National Education Association. *The Questionnaire*. Research Bulletin, Vol. VIII, No. 1. January 1930.
- RUCH, GILES M., AND SEGEL, DAVID. *Minimum Essentials of the Individual Inventory in Guidance*. Vocational Division Bulletin, No. 202, Occupational Information and Guidance Series, No. 2. Washington, D. C.: U. S. Government Printing Office, 1940.
- SEGEL, DAVID, AND OTHERS. *Handbook of Cumulative Records*. U. S. Office of Education Bulletin, 1944, No. 5. Washington, D. C.: U. S. Government Printing Office, 1944.
- SHANE, HAROLD G., AND MCSWAIN, E. T. *Evaluation and the Elementary Curriculum*. New York: Henry Holt and Co., 1951. Chapter 3.
- SYMONDS, PERCIVAL M. *Diagnosing Personality and Conduct*. New York: D. Appleton-Century Co., Inc., 1931. Chapters 4, 12.
- TABA, HILDA. "Planning and Administering the Evaluation Program." *Appraising and Recording Student Progress*. New York: Harper and Brothers, 1942. Chapter 8.
- TABA, HILDA, AND MCGUIRE, CHRISTINE. "Evaluation of Social Sensitivity." *Appraising and Recording Student Progress*. New York: Harper and Brothers, 1942. Chapter 3.
- TABA, HILDA, AND OTHERS. *Diagnosing Human Relations Needs*. Studies in Intergroup Relations. Washington, D. C.: American Council on Education, 1951.
- THUT, I. N., AND GERBERICH, J. RAYMOND. *Foundations of Method for Secondary Schools*. New York: McGraw-Hill Book Co., Inc., 1949. Chapters 14, 16.
- TRAXLER, ARTHUR E. "Cumulative Records: Their Nature and Uses." *Educational and Psychological Measurement*, 1:323-40; October 1941.
- WRIGHTSTONE, J. WAYNE. "Evaluating Achievement." *Childhood Education*, 39:561-82; April 1946.
- WRIGHTSTONE, J. WAYNE. "Evaluation." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 403-7.
- WRIGHTSTONE, J. WAYNE. "Trends in Evaluation." *Educational Leadership*, 8:91-95; November 1950.
- WRINKLE, WILLIAM L. *Improving Marking and Reporting Practices*. New York: Rinehart and Co., Inc., 1947.



## *Using Intelligence and Aptitude Tests*

THE ASPECTS of intelligence, intelligence testing, and the use of intelligence test results that are given major attention in this chapter are as follows:

- A. Definitions of intelligence.
- B. Theories concerning the nature of intelligence.
- C. Individual and group tests of intelligence.
- D. Specific and group-factor intelligence tests.
- E. Performance tests of intelligence.
- F. Scores derived from intelligence tests.
- G. Distribution of intelligence.
- H. Procedures in intelligence testing.
- I. Classroom uses of intelligence test results.

It is important for the student to be conversant with the nature of intelligence and with techniques for its measurement. It is also important that he be able to obtain and use at least the major types of derived scores in furnishing guidance of various types to his pupils. This chapter discusses the theory and measurement of intelligence and the applied aspects of intelligence and intelligence testing.

Workers in the field of mental abilities are far from agreement both on the correct terminology to use in discussing mental abilities and on the exact nature of the ability or abilities to which the terms apply. It is therefore very difficult to prepare a brief treatment of

intelligence and intelligence testing. The discussions of intelligence in this chapter are based on what the authors believe to be the best modern terminology in this field. The reader will doubtless encounter instances, however, in which test titles and references will not be completely in harmony with the usage to be followed.

## 1 NATURE OF INTELLIGENCE

The exact nature of the combination of abilities known as intelligence is not well understood. However, it is definitely known that individuals differ widely in the amount, and perhaps the quality, of it they possess, and that within limits it can be measured.

### Definitions of general intelligence

Many definitions of intelligence have been given. The following list presents some of the ones that are most commonly quoted:<sup>1</sup>

Colvin: "An individual possesses intelligence in so far as he has learned, or can learn to adjust himself to his environment."

Dearborn: "... the capacity to learn or to profit by experience. . ."

Henmon: "Intelligence . . . involves two factors—the capacity for knowledge and knowledge possessed."

Pintner: "I have always thought of intelligence as the ability of the individual to adapt himself adequately to relatively new situations in life."

Terman: "An individual is intelligent in proportion as he is able to carry on abstract thinking."

Thorndike: "We may . . . define intellect, in general, as the power of good responses from the point of view of truth or fact."

Woodrow: "It is an acquiring-capacity."

Additional definitions taken from Freeman<sup>2</sup> are:

Binet: "... the tendency of thought to take and maintain a definite direction, the capacity to make adaptations for the purpose of attaining the desired end, and the power of self-criticism."

Burt: "... the power of readjustment to relatively novel situations. . ."

Stern: "... the general mental adaptability to new problems and conditions of life."

<sup>1</sup> Symposium, "Intelligence and Its Measurement." *Journal of Educational Psychology*, 12:123-47, 195-216; March and April 1921.

<sup>2</sup> Frank N. Freeman, *Mental Tests: Their History, Principles and Applications*, Revised edition. Houghton Mifflin Co., Boston, 1939. p. 248.



The above definitions seem to fall into at least three patterns: (1) the rather formal definitions stressing mainly what have been called the higher mental powers, (2) the definitions emphasizing ability to learn, and (3) the definitions placing major emphasis upon adaptability. It is felt that the last type of definition particularly, by which intelligence is conceived as *the capacity or power of the individual to adapt himself* to his environment and to new situations, is the most meaningful for the purposes of the teacher. However, the fact that capacity to learn and ability to think in abstract terms are both evidences of intelligence should not be overlooked.

Freeman<sup>3</sup> listed three concepts of intelligence—the organic, the social, and the psychological or behavioristic. He considered that the third is the only one that is of direct concern to intelligence testers and called the others factors of intelligence. The psychological or behavioristic concept accepts as intelligence the types of behavior that are measured by intelligence tests. Intelligence has been defined as “that which intelligence tests measure.” This definition is in line with Freeman’s psychological or behavioristic concept. The definition has meaning, for it implies that intelligence, although it has not yet been adequately defined or delimited, conditions the individual’s behavior and that it is, therefore, through observation and measurement of his behavior that his intelligence can be estimated.

Stoddard also approached intelligence operationally, although he stated that the elements have not been represented, unless accidentally, in existing tests, and defined it as:

the ability to undertake activities that are characterized by (1) difficulty, (2) complexity, (3) abstractness, (4) economy, (5) adaptiveness to a goal, (6) social value, and (7) the emergence of originals, and to maintain such activities under conditions that demand a concentration of energy and a resistance to emotional forces.<sup>4</sup>

## Theories concerning intelligence

Theories concerning the nature of ability go back as far as pronouncements of the early philosophers. However, only three of the most important theories of the last century are presented here. Two

<sup>3</sup> Frank N. Freeman, “The Meaning of Intelligence,” *Intelligence: Its Nature and Nurture*, Thirty-Ninth Yearbook of the National Society for the Study of Education, Part I. Public School Publishing Co., Bloomington, Ill., 1940. p. 11-20.

<sup>4</sup> George D. Stoddard, *The Meaning of Intelligence*. Macmillan Co., New York, 1943. p. 4.

of them are important to the user of intelligence tests because of the manner in which they have modified and are now modifying teaching and testing practices.

*The faculty theory.* According to the faculty theory, intelligence consists of a number of relatively independent and largely correlated and specialized abilities of various types, such as memory, imagination, honesty, and language ability, to name only a few. The closely related theory of formal discipline maintained that these faculties could be developed individually by means of general mental exercise. However, when the theory of formal discipline was disproved and the transfer of training concept directed attention to the fact that such faculties as those named above are neither psychological entities nor subject to general training, the faculty theory was forced into the discard as an explanation of mental abilities.

*The two-factor theory.* Spearman first presented his two-factor theory in 1904.<sup>5</sup> He proposed a general factor, or *g*, which enters into all types of performance, and many specific factors, called *s*, which combine with *g* to determine total activity. Basing his theory on technical statistical treatments of data, Spearman later added a third type of factor, called *group factors*, which represent the overlap among *s* factors.<sup>6</sup> Thus, according to his theory, a *g* or general factor, which might be called energy, *group factors*, such as number ability and mechanical ability, and many *s* or specific factors constitute ability.

*The multi-factor theory.* Spearman's work may be considered the forerunner of the present *factor analysis* approach to the nature of mental ability. Among the factor analysts is Thurstone, who isolated the seven factors of perceptual, number, verbal, spatial, memory, inductive reasoning, and deductive reasoning,<sup>7</sup> which he called primary mental abilities. These primary abilities might appear on the surface to relate closely to the "faculties" of the early psychologies, but the factors emerging from the work of Thurstone and other exponents of the *multi-factor* theory not only are substantiated by correlational relationships but also appear to have sound psychological evidence to support their existence.

<sup>5</sup> C. Spearman, "'General Intelligence' Objectively Determined and Measured." *American Journal of Psychology*, 15:201-93; 1904.

<sup>6</sup> C. Spearman, *The Abilities of Man*. Macmillan Co., New York, 1927. p. 82.

<sup>7</sup> Louis L. Thurstone, *Primary Mental Abilities*. Psychometric Monograph Series, No. 1. University of Chicago Press, Chicago, 1938.



## 2 MEASUREMENT OF INTELLIGENCE

### Indirect measurement of intelligence

For practical purposes, intelligence has been defined in a preceding section of this chapter as the power to learn or to adapt to new situations. These definitions perhaps suggest that this type of ability is subject to evaluation in a rather direct manner. Such is not the case, however, for ability to learn can only be inferred from the fact that learning has occurred in a test situation. Since intelligence itself cannot be measured, test makers can only measure the performance of tasks the successful completion of which is generally believed to be dependent upon intelligence. The value of the intelligence test lies in the fact that it affords an objective basis for this inference. It samples widely from the fields of learning resulting from experiences assumed to be common to all persons subjected to the test. The pupil's capacity to learn or to adapt to new situations is determined by summing up his reactions to the items of the test.

There is apparently no way of determining very precisely which particular fields of human interest or ability should be sampled in the attempt to secure this cross section of mental activity. It is important that the sampling be sufficiently diverse and representative to permit the securing of an estimate in the nature of an average that will not penalize a person because he may not have had this or that specific experience. Briefly, the measure or average obtained from a test that does sample representative reactions is taken to be indicative of one's ability to learn or of one's adaptability. Roughly, it is assumed that what an individual has learned is indicative of his potentialities for learning. Differences in intelligence test scores are probably sufficiently accurate, rough as they are, to indicate such differences in mental ability.

### Factual and skill content of intelligence tests

It has been contended, and not without justification, that intelligence tests do not differ appreciably from achievement tests, inasmuch as both are founded upon the measurement of knowledges and skills that have largely been learned. Obviously, a test of ability to learn must have some type of content. Intelligence tests admittedly

contain factual and skill materials. Such tests attempt to measure abilities to see relationships, to draw reasoned inferences, to manipulate, to compare, to contrast, and otherwise to handle materials which themselves are so commonly known and at such low difficulty levels that all persons who have had any but the most exceptional environmental backgrounds should know the necessary facts and have the necessary skills for understanding and taking, although not necessarily for succeeding upon, the tests. To contend that intelligence tests have been completely successful in eliminating the significance of the factual and skill content would be foolhardy and contrary to available evidence.

A few intelligence tests contain vocabulary sections requiring considerable knowledge of word meanings for successful performance. Several tests also directly measure knowledges in widely studied areas. The justification for the inclusion of factual items in an intelligence test is that opportunities are supposed to be similar for all persons experiencing a normal environment to learn such facts and that the degree to which different persons do so is partial evidence concerning their intellectual levels. More frequently, however, intelligence tests attempt, but with varying degrees of success, to minimize the influence of environment upon an individual's test performance.

Kelley<sup>8</sup> stated that general intelligence tests and achievement tests overlap to the degree indicated by a correlation coefficient of .90. In general, coefficients of .40 to .60 are found between tested intelligence and academic achievement, but higher degrees of relationship are sometimes found. When such correlations approach .70 or .80, the intelligence test is looked upon with suspicion by some and may be considered a general scholastic achievement test rather than an intelligence test.<sup>9</sup>

Cattell, believing that general intelligence tests measure acquired knowledges and skills to a considerable degree and also that they frequently test abilities of too specific a nature, devised a culture-free test.<sup>10</sup> The test items, largely pictorial rather than verbal, were chosen to measure abilities to run pencil mazes, to build up series,

<sup>8</sup> Truman L. Kelley, *Interpretation of Educational Measurements*. World Book Co., Yonkers, N. Y., 1927. p. 208.

<sup>9</sup> Paul L. Boynton, "Intelligence." *Encyclopedia of Educational Research*. Macmillan Co., New York, 1941. p. 630.

<sup>10</sup> Raymond B. Cattell, "A Culture-Free Intelligence Test I." *Journal of Educational Psychology*, 31:161-79; March 1940.



to classify, and to determine relationships of varying degrees of complexity. The content was so selected as to be largely independent of acquired or learned *meaning*, so that the test presumably can be given with equal fairness to persons reared in any civilized society and even, by pantomime, to primitive peoples.

The teacher should probably admit that intelligence tests in varying degrees test factual knowledges and skills which not all pupils have had equal opportunities to learn, but he is probably justified in the belief that they are at a minimum in at least the better tests and that the environments of pupils attending the typical school are sufficiently similar that all have had approximately equal opportunities to acquire such facts and skills as are included in the tests.

### 3 GENERAL INTELLIGENCE TESTS

General intelligence tests, both individual and group, are discussed and illustrated below so that the student may obtain a more complete understanding of the characteristics and representative content of these important instruments for the measurement of general mental ability.

#### Individual scales of general intelligence

Individual intelligence examinations constitute the most accurate devices for the measurement of intelligence. The length of the test, the wide variety of reactions called for, the fact that the subject receives his instructions personally from the examiner, the fact that the examiner is afforded an opportunity to observe each reaction made by the subject, and the careful standardization of procedures for administering the test and scoring the subject's reactions all contribute to the high degree of accuracy. The full time of an examiner is required for each pupil tested. The examiner must be a person who is more capable and efficient in test administration than is the typical teacher. Furthermore, he must be one who has had extensive training and experience in giving individual intelligence tests.

Individual intelligence tests are largely patterned upon the *Binet-Simon* tests brought out in France from 1905 to 1911. American

adaptations and revisions were published by Goddard in 1911, Kuhlmann in 1912, Terman in 1916, Herring in 1922, and Terman and Merrill in 1937. The Terman and Merrill *New Revised Stanford-Binet Tests of Intelligence* is today, as was its 1916 predecessor, the best known and most widely used individual test of general intelligence in America.

The general procedure in administering the *New Stanford-Binet* is quite representative of that of the other revisions mentioned. The type of performance tested varies considerably with the different exercises. These test elements are presented to the child by means of spoken directions. The test should be given in a quiet room where there is freedom from distraction. A friendly attitude between examiner and subject should be maintained. The examiner is expected to make sure that the subject understands what is to be done, and in all cases the burden of proof is with the examiner to show that the subject has responded in a way that is representative of his ability.

After rapport has been established, i.e., the child has been put at ease, the examiner starts the test with materials at a scale level on which the subject is likely to succeed with some effort. If he is successful on all tests at this level, the examiner, assuming that he could pass all tests at lower levels, passes on to the higher levels and continues on through the scale until the subject fails all tests at one age level. In effect the child has been tested over the entire scale, for his success on all tests at one age level makes almost certain that he could pass all tests at lower levels and his failure on all tests of another, and higher, age level indicates with essential certainty that he could go no higher on the scale.<sup>11</sup> The child's mental age is determined by giving him credit for the number of years below the level on which he passes all tests and adding to this amount the years and months of credit assigned to the higher level tests he succeeds in passing.

It is not feasible here to reproduce more than a few sample test elements, but the two following samples from the *New Stanford-Binet*, chosen from those most easy to reproduce in limited space, will give the student some idea of the nature of the test.

<sup>11</sup> Lewis M. Terman and Maud A. Merrill, *Measuring Intelligence*. Houghton Mifflin Co., Boston, 1937. p. 63.



Year III-6, Form L, Test 3, Comparison of Sticks<sup>12</sup>*Comparison of Sticks*

*Material:* Match sticks, cut to 2-inch and 2½-inch lengths.

*Procedure:* Place the two sticks on the table before the child in the positions indicated below and about an inch apart. Say, "*Which stick is longer?*" "*Put your finger on the long one.*" Give three trials, alternating the relative positions of the long and the short sticks. In case one of the first three trials is failed, give three additional trials, continuing to alternate the positions of the sticks.

(a) \_\_\_\_\_ (b) \_\_\_\_\_ (c) \_\_\_\_\_  
 \_\_\_\_\_

Score: 3 of 3 or 5 of 6.

Superior Adult I, Form L, Test 2, Enclosed Box Problem<sup>13</sup>*Enclosed Box Problem*

*Material:* Any small cardboard box.

*Procedure:* Show S. a box and say:

(a) "*Listen carefully. Let's suppose that this box has 2 smaller boxes inside it, and each one of the smaller boxes contains a little tiny box. How many boxes are there altogether, counting the big one?*"

(b) "*Now let's suppose that this box has 2 smaller boxes inside it and that each of the smaller boxes contains 2 tiny boxes. How many altogether?*"

(c) "*Now suppose that this box has 3 smaller boxes inside it and that each of the smaller boxes contains 3 tiny boxes. How many boxes are there altogether?*"

(d) "*Now suppose that this box has 4 smaller boxes inside it and that each of the smaller boxes contains 4 tiny boxes. How many are there altogether?*"

Score: 3 plus.

The lists of test titles<sup>14</sup> at several age levels of the Form L *Stanford-Binet* between Year II and the Superior Adult III, which represent the bottom and top of the scale, will indicate the variety of abilities tested, the scalar arrangement of tests from easy to diffi-

<sup>12</sup> *Ibid.* p. 84.

<sup>13</sup> *Ibid.* p. 125.

<sup>14</sup> *Ibid.* p. 75-132.

cult, and the duplication at different age levels of similar types of test situations at varying levels of difficulty.

#### Year II

Three-Hole Form Board  
Identifying Objects by Name  
Identifying Parts of the Body  
Block Building: Tower  
Picture Vocabulary  
Word Combinations

#### Year V

Picture Completion: Man  
Paper Folding: Triangle  
Definitions  
Copying a Square  
Memory for Sentences II  
Counting Four Objects

#### Year VIII

Vocabulary  
Memory for Stories: The Wet Fall  
Verbal Absurdities I  
Similarities and Differences  
Comprehension IV  
Memory for Sentences III

#### Year XII

Vocabulary  
Verbal Absurdities II  
Response to Pictures II  
Repeating 5 Digits Reversed  
Abstract Words II  
Minkus Completion

#### Average Adult

Vocabulary  
Codes  
Differences between Abstract Words  
Arithmetical Reasoning  
Proverbs I  
Ingenuity  
Memory for Sentences V  
Reconciliation of Opposites



## Superior Adult III

Vocabulary

Orientation: Direction II

Opposite Analogies II

Paper Cutting II

Reasoning

Repeating 9 Digits

## Group tests of general intelligence

Group intelligence tests originated in America during World War I. The *Army Alpha* and *Army Beta* tests, the latter really a performance scale, were developed for use in selecting army recruits for officers' training and for other positions requiring high intelligence. Shortly after the war, Otis, Terman, and others brought out group tests devised for use in the schools, and many such tests were published between 1918 and 1925. Approximately ten years then elapsed during which few new group tests of general intelligence made their appearance. The *Army General Classification Test*, the *Aviation Cadet Qualifying Examination*, and the *Navy General Classification Test* of World War II are representative of the recent counterparts of *Army Alpha*. A number of revisions of earlier tests and of new tests for civilian use have also been published since 1935.

Space limitations prevent the use of illustrations from more than a few intelligence tests and permit only a brief treatment of the testing techniques used. No attempt is made to furnish descriptions of any of the group tests of general intelligence. Instead, sample items of various types representative of testing techniques are shown and briefly commented upon. The only way by which the student can become truly familiar with intelligence tests is by examination and actual use of them.

The accompanying illustration of two of the *Kuhlmann-Anderson Intelligence Tests* shows parts that measure knowledge of the alphabet and ability to follow directions of various types. These are among the higher level tests of a series of 39 that are divided into nine booklets for use from the first grade to maturity. Bases are provided for interpreting the results in terms of the mental age (*MA*), and intelligence quotient (*IQ*) and also in terms of the personal constant (*PC*), all of which are discussed later in this chapter.

Excerpts from Kuhlmann-Anderson Intelligence Tests<sup>15</sup>

## TEST 27

A B C D E F G H I J K L M N O P Q R S T U V W X Y Z

## EXAMPLES:

The third letter of the alphabet is . . . . .

The second letter before the sixth letter is . . . . .

1. The fifth letter of the alphabet is . . . . . 1
2. The second letter before the last letter is . . . . . 2
3. The third letter before M is . . . . . 3
4. The letter midway between H and N is . . . . . 4
5. The second letter after the fourth letter is . . . . . 5

## TEST 35

Draw a line under the middle one of these three numbers: 3 8 9.

Write here ..... a word meaning the opposite of *good*.

Draw a line through the middle letter in the longer of these two words: Revenge, Assert. Write here ..... a word of five letters meaning the opposite of *slow*. Write here .....

a word which rhymes with *hay* and means a part of a week.

Draw a line after each of these two letters A B making the first line half as long as the second. Think what year this is, then write here ..... the digits in the reverse order, the one which belongs last coming first. Cross out one digit in each of these numbers which does not appear in the other number: 43689, 64378.

<sup>15</sup> F. Kuhlmann and Rose G. Anderson, *Kuhlmann-Anderson Intelligence Tests*, Sixth edition. Published by Personnel Press, Inc., 1952.



## 4 SPECIFIC INTELLIGENCE TESTS

The two types of specific intelligence tests—aptitude and readiness—differ primarily on the age and maturity levels of the pupils to whom they are given. Whereas aptitude tests presuppose some ability to read and compute, readiness tests assume that such skills have not yet been acquired. Aptitude tests first made their appearance nearly forty years ago, but readiness tests have been in use mainly since the early thirties.

### Aptitude tests

Aptitude tests are now available for a number of areas of performance, such as those involved in various occupations in the trades and industry, various broad areas of performance commonly dealt with in the school, and various narrow areas of performance largely unique to the school. The various types of aptitude tests largely possess in common the characteristic of testing the individual's potentialities in terms of the specific abilities resulting from inheritance and general experience but of disregarding the abilities resulting from specific training or education. Thus aptitude tests parallel intelligence tests, although they are narrower in scope.

Teachers and school officers, aside from those engaged in vocational guidance and placement, are more concerned with aptitudes for school subjects and fields of study than with occupational areas. Therefore, occupational aptitude tests, sometimes called trade tests, will not be discussed intensively in this volume but will receive treatment only insofar as some of them are useful in the schools.

Among the first tests of aptitude to be developed primarily for school use were several for mechanical, musical, artistic, and clerical abilities. In the academic areas of English, foreign languages, mathematics, and the sciences, the *Iowa Placement Examinations, Aptitude Series*, published in 1925, appear to be the pioneer instruments. These tests, primarily useful at the college level, were followed by other aptitude tests for algebra and geometry, English, the foreign languages, mathematics, and the sciences for secondary-school use. The accompanying excerpt from the *Iowa Algebra Aptitude Test* illustrates the number series type of item rather common to aptitude tests in mathematics. It is apparent that some persons who could

perform the necessary arithmetical operations for answering item 5, for example, would not do so because they failed to discover the "pattern" of the number series.

The variety of areas of behavior served by aptitude tests makes impracticable a comprehensive discussion of such instruments here. They will receive consideration in Chapters 15 to 21 by subject fields, in parallel with prognostic tests, which, although frequently measuring the results of training, have somewhat similar uses. Aside from tests in the music and art fields, aptitude tests are devised almost exclusively for use at the high-school and college levels.

### Excerpt from Iowa Algebra Aptitude Test <sup>16</sup>

#### Part 3. NUMERICAL SERIES

Time allowance—12 minutes.

**Directions:** Each of the following number series is made up according to some rule. Addition, subtraction, multiplication, and division, and various combinations of these processes are used in forming the different series. Discover the rule for each example, decide what the next term would be, and write it on the blank line following the series. Then place a cross (X) in the circle directly over the answer that agrees with yours. If no answer agrees with yours place the X in the circle over "Not Given." You will receive no credit for a correct answer unless it is marked in the correct answer space. The sample is answered correctly.

credit for a correct answer unless it is marked in the correct answer space.											Answers			
Sample	1	2	3	4	5	6	7	8			7	<input checked="" type="radio"/>	9	Not Given
											ANSWERS			
1.	2		4		6		8	10	_____	1.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Not Given
2.	9		8		7		6	5	_____	2.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Not Given
3.	1		1		5		5	9	9 _____	3.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Not Given
4.	2		4		8		16	32	_____	4.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Not Given
5.	5		8		11		14	17	_____	5.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	Not Given

### Readiness tests

Readiness tests, found primarily in reading and arithmetic, are largely tests of specific intelligence, for they measure the results of inheritance and general training rather than of direct instruction. As readiness tests imply by their general designation, they measure readiness to undertake a new type of activity that is dependent upon the maturation of various physical and mental abilities. They may in one sense be considered as aptitude tests at the elementary- and even the primary-school levels, where they almost entirely occur.

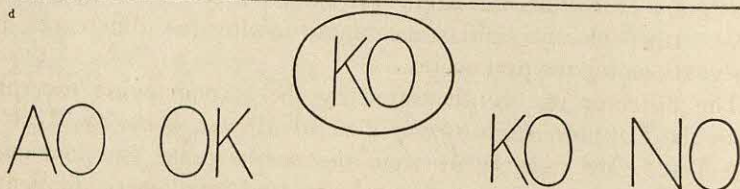
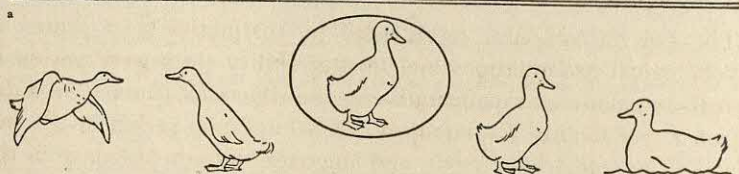
<sup>16</sup> H. A. Greene and A. H. Piper, *Iowa Algebra Aptitude Test*, Revised edition. Published by Bureau of Educational Research and Service, University of Iowa, 1942.



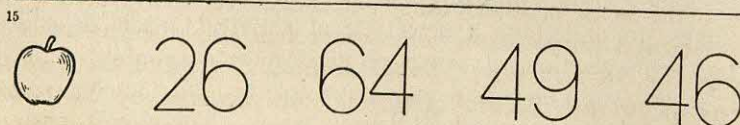
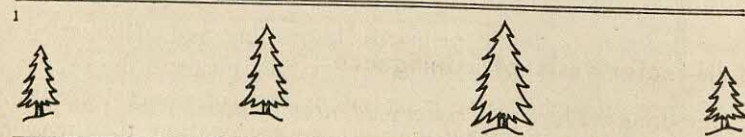
Tests of this type are usually restricted in applicability to a particular subject field. However, the *Metropolitan Readiness Tests* are devised for determining the readiness of a child to learn first-grade skills of all types, and consequently are briefly discussed and illustrated here. The six parts of the test seem to measure the types of abilities used primarily in reading and number work. Tests 4 and 5, for which the instructions are given orally by the examiner and which require few skills in pencil manipulation of any complexity, measure respectively ability in visual perception and knowledge of number.

### Excerpts from Metropolitan Readiness Tests <sup>17</sup>

#### TEST 4. MATCHING



#### TEST 5. NUMBERS



<sup>17</sup> Gertrude H. Hildreth and Nellie L. Griffiths, *Metropolitan Readiness Tests*, Form S. Published by World Book Co., 1950.

## 5 GROUP-FACTOR TESTS OF INTELLIGENCE

The factor analysis movement gave rise some twelve or fifteen years ago to the first use of group factors of intelligence in testing practice. The bi-factor tests made their appearance first, and it has been only during the past five or so years that the multi-factor tests have passed their experimental stage. Both types of group-factor tests may be considered to measure intellectual abilities less broad than general intelligence but in major respects broader than the areas measured by specific intelligence tests.

### Bi-factor tests of intelligence

The two factors first represented by distinctive part scores in psychological examinations and mental ability tests were variously termed linguistic and quantitative in the *American Council on Education Psychological Examination*, verbal and non-verbal in the two *Pintner General Ability Tests*, and language and non-language in the *California Test of Mental Maturity*. These pairs of scores appear to have rather closely similar meanings despite the differences in designations for the part scores.

The bi-factor test is illustrated by the accompanying excerpts from the *California Short-Form Test of Mental Maturity*. Test 2 and Test 7 are respectively from the non-language and language portions of the instrument. Mental ages and intelligence quotients can be obtained separately for these two major factors as well as for general intelligence.

### Multi-factor tests of intelligence































Thurstone's *Tests of Primary Mental Abilities* were pioneers in the multi-factor types of tests, but they were used primarily for experimental purposes for some years following their issuance in 1938. The *SRA Primary Mental Abilities Tests* and the *Chicago Tests of Primary Mental Abilities*, both prepared by Thurstone, have been in use for several years in the measurement of such factors of intelligence as verbal-meaning, space, reasoning, memory, number, and word-fluency at the age 11 to 17 level. Comparable tests by Thurstone for younger children specify similar lists of factors.



Excerpts from California Short-Form Test of Mental Maturity<sup>18</sup>

**DIRECTIONS:** In each row find the drawing that is a different view of the first drawing. Mark its number as you are told.

**TEST 2**

<p><b>C</b></p>  <p>1  2  3  4 </p> <p>21</p>  <p>1  2  3  4 </p> <p>22</p>  <p>1  2  3  4 </p>	<p>28</p>  <p>1  2  3  4 </p> <p>29</p>  <p>1  2  3  4 </p> <p>30</p>  <p>1  2  3  4 </p>
--	--

**DIRECTIONS:** Mark as you are told the number of the word that means the same or about the same as the first word.

**TEST 7.**

- H. blossom 1 tree 2 vine  
3 flower 4 garden — H
96. strange 1 real 2 tell  
3 certain 4 unknown — 96
97. reply 1 news 2 answer  
3 note 4 open — 97
98. liberty 1 benefit 2 seize  
3 freedom 4 aid — 98
99. assist 1 consent 2 help  
3 agree 4 overlook — 99

120. invariably 1 probably 2 seldom  
3 always 4 motionless — 120
121. detect 1 remove 2 discover  
3 overtake 4 apply — 121
122. reluctantly 1 gladly 2 instantly  
3 certainly 4 unwillingly — 122
123. inefficient 1 unruly 2 prudent  
3 incompetent 4 inevitable — 123
124. facetious 1 active 2 fragile  
3 humorous 4 inventive — 124
125. ambiguous 1 hard 2 doubtful  
3 responsible 4 confident — 125

The accompanying illustration from four of the *Differential Aptitude Tests* show some of the techniques used in multi-factor tests. The verbal reasoning test employs an analogy type of item in which a numbered response is used to fill the first blank and a lettered response is required to complete the analogy. In the abstract reasoning test the appropriate lettered response is selected to carry on the progression established in the four left-hand figures. The problem in the space relations test is to select the lettered response that represents the three-dimensional figure resulting when the left-hand figure is folded and assembled. In the mechanical reasoning test the nature of the problem is self-evident.

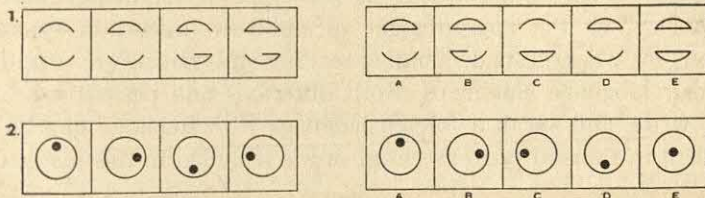
<sup>18</sup> Elizabeth T. Sullivan, Willis W. Clark, and Ernest W. Tiegs, *California Short-Form Test of Mental Maturity*, Intermediate. Published by California Test Bureau, 1950.

Excerpts from Differential Aptitude Tests <sup>19</sup>

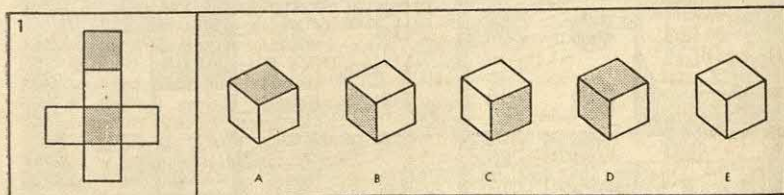
VERBAL REASONING

1. ....is to street as rd. is to....
- |         |           |        |         |
|---------|-----------|--------|---------|
| 1. lo.  | 2. ma.    | 3. st. | 4. aw.  |
| A. city | B. France | C. end | D. road |
2. ....is to cavalry as foot is to....
- |          |             |           |             |
|----------|-------------|-----------|-------------|
| 1. horse | 2. cemetery | 3. votary | 4. hiding   |
| A. yard  | B. travel   | C. armory | D. infantry |

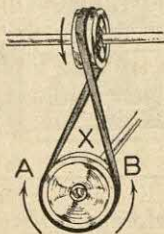
ABSTRACT REASONING



SPACE RELATIONS



MECHANICAL REASONING



2

When the top pulley turns in the direction shown, which way will the lower pulley turn?  
(If either, mark C.)

<sup>19</sup> George K. Bennett, Harold G. Seashore, and Alexander G. Wesman, *Differential Aptitude Tests*, Form A. Copyright by Psychological Corporation, 1947.





Two types of performance tests may be distinguished—those requiring the use of a pencil for marking, but not for writing, and those requiring manipulations of various items of testing equipment.

The *Army Beta*, and such revisions as the *Kellogg-Morton Revised Beta Examination* for use with adults who cannot read, write, or perhaps even understand English, illustrate the first type. Directions are given by pantomime, and the subjects respond by tracing mazes, indicating whether groups of numbers are alike or unlike, and supplying missing elements in pictures.

The second type of test, requiring manipulation of apparatus, depends largely upon form boards that are not unlike jigsaw puzzles. The accompanying reproduction of the tests comprising the *Pintner-Paterson "Long" Performance Scale* shows the general nature of form boards used in the measurement of mental ability. Directions are usually given orally by the examiner. The pupil's success is measured by time, errors, moves, and other evidences of success or failure. Fifteen separate tests, each of which nets a mental age score, are included in the scale. The median of these mental ages is taken as the pupil's mental ability measure.

## 7 DERIVED RESULTS OF INTELLIGENCE TESTING

A raw score from a test has little or no meaning unless it can be compared in some manner with other similarly obtained and comparable raw scores. This general principle applies to intelligence tests as well as to achievement tests. Therefore, it is important that the teacher know the meaning of, and the method of obtaining, the most common types of derived measures used in the interpretation of intelligence test results. As the methods of obtaining the most important of the derived scores discussed below are given fully in Chapter 13, only general meanings are treated here.

### Mental age (MA)

Terman defined mental age as "that degree of general mental ability which is possessed by the average child of corresponding chronological age," and as "an index of absolute mental level" indicating "the level of development which a child has reached at a given time."<sup>21</sup> For example, a child has a mental age of ten years

<sup>21</sup> Lewis M. Terman, *The Intelligence of School Children*. Houghton Mifflin Co., Boston, 1916. p. 7-8.



if his level of mental development is equal to that of the normal child of exactly ten years. Thus if a representative group of pupils, all of whom are ten years of age, makes an average score of 45 on an intelligence test that is being standardized, any pupil who subsequently takes this test and earns a score of 45 is said to have a mental age of ten years. An average score for each age group is established in the same manner.

The mental age (*MA*) is a measure of *mental level* or of *mental maturity* of the individual. Taken alone it tells nothing of how relatively bright or dull the child may be, but it does give an indication of the level of ability at which the child potentially can work. For example, information to the effect that a certain child has a mental age of 7-6 does not enable a person to judge whether the child is bright, average, or dull. It is only when he knows or at least can estimate the child's chronological age that he can draw conclusions concerning the child's brightness.

The mental age should probably be considered a specific rather than a general concept. That is, a child does not have just one mental age at a given time; he has many.<sup>22</sup> His mental age, then, depends upon the particular test or tests by which it has been determined, and such tests may be specific intelligence and group-factor tests as well as general intelligence tests, although the latter are the tests in the areas of mental ability most commonly providing mental age norms.

### Intelligence quotient (*IQ*)

When the chronological age (*CA*), i.e., life age in years and months, is known for a pupil, and his mental age (*MA*) has been determined from his score on an intelligence test, his intelligence quotient (*IQ*) can be computed. The intelligence quotient is a simple method of expressing the relationship between a pupil's mental age and his chronological age. To obtain the *IQ*, a child's mental age (in months) is divided by his chronological age (in months), the result is multiplied by 100 to remove the decimal point, and the whole number nearest to the result is taken as his intelligence quotient. The formula is:

$$IQ = 100 \frac{MA}{CA}.$$

<sup>22</sup> Terman and Merrill, *op. cit.* p. 25.

If this formula is applied for a child who has a mental age of twelve years six months (150 months) when he is ten years five months (125 months) of age chronologically, the following is the result:

$$IQ = 100 \frac{12-6}{10-5} = 100 \frac{150}{125} = 120.$$

The intelligence quotient is a measure of the pupil's relative *brightness*. If it is assumed that a typical child grows in mentality at the same rate as he ages chronologically, it then appears that children who have *IQs* over 100 are above average and children who have *IQs* below 100 are below average. This is in harmony with the usual indication of normal intelligence as being represented by *IQs* between 90 and 110, for people of normal intelligence center around but are not necessarily exactly at the average of intelligence. However, as this concept of the average is applicable only in terms of the population as a whole and as very few pupil groups are average in this sense, the teacher should not generalize this statement and make it apply to pupil groups in the school. The *IQ* alone tells nothing about the level of work of which a child is capable, for two children of age six and age twelve might both have *IQs* of 110 and yet the younger child would be entirely incapable at that time of types of performance commonplace to the older child.

*The mental growth curve.* The curve of mental growth has long been under scrutiny and has been subjected directly and indirectly to many research studies by psychologists. However, no completely satisfactory unit of mental growth has yet been found. This fact, which results from the lack of an absolute zero point of intelligence, from the lack of a simple and constant mental growth unit, and for other technical reasons, gives rise to a major problem in the measurement of intelligence for persons beyond their late-middle teens in chronological age. In practice, intelligence tests handle this problem in various ways, but a common method is to use the individual's actual chronological age in computing the *IQ* until he attains the age of fourteen to eighteen and from that point to assume for purposes of computing his intelligence quotient that his chronological age remains constant for the remainder of his life. The justification for doing so is found in the shape of the mental growth curve. Progressing upward very rapidly during early life, and slowing down somewhat during childhood and the early teens, it flattens out to almost a horizontal



line by the age of sixteen or so. Although Thorndike presented evidence to show that mental growth continues into the early twenties,<sup>23</sup> the annual increments or additions beyond the age of sixteen are quite small.

Constancy of the IQ. A heated controversy over the constancy of the intelligence quotient has been waged during the last fifteen years. Although it has been recognized for many years that the *IQ* obtained by the use of the best modern tests fluctuates within limits because the tests are not perfectly reliable, and that major environmental changes for an individual may well be reflected in his *IQ*, rather startling evidence was presented some twenty years ago<sup>24</sup> to show average gains of twenty *IQ* points for 600 children who had attended preschool for four years. Later and more startling evidence<sup>25</sup> showed that children of dull parentage who were placed in foster homes shortly after birth had mean intelligence quotients of 116 when they were tested a few years later. These and other studies support the belief that the intelligence quotient is significantly influenced by very favorable environments.

Although such findings have not been uniformly obtained by experimenters,<sup>26</sup> they are supported by other types of experimental evidence revealing at least the possibility of marked changes in intelligence quotients as the result of improved environments.<sup>27</sup> Stoddard summed up the evidence on inconstancy of the *IQ*<sup>28</sup> and pointed out Binet's expression of the belief<sup>29</sup> that the *IQ* is subject to improvement under desirable conditions of stimulation.

<sup>23</sup> Edward L. Thorndike, Elsie O. Bregman, and Ella Woodyard, *Adult Learning*. Macmillan Co., New York, 1928. p. 127.

<sup>24</sup> Beth L. Wellman, "The Effect of Pre-School Attendance on the IQ." *Journal of Experimental Education*, 1:48-69; September 1932.

<sup>25</sup> Harold M. Skeels, "Mental Development of Children in Foster Homes." *Journal of Consulting Psychology*, 2:33-43; March-April 1938.

<sup>26</sup> Florence L. Goodenough and Katharine M. Maurer, "The Mental Development of Nursery-School Children Compared with That of Non-Nursery-School Children." *Intelligence: Its Nature and Nurture*, Thirty-Ninth Yearbook of the National Society for the Study of Education, Part II. Public School Publishing Co., Bloomington, Ill.; 1940. p. 161-78.

<sup>27</sup> Percival M. Symonds, "Psychological Tests and Their Uses: Review and Preview." *Review of Educational Research*, 8:217-20; June 1938.

<sup>28</sup> George D. Stoddard, "The IQ: Its Ups and Downs." *Educational Record*, 20:44-57, Supplement No. 12; January 1939.

<sup>29</sup> Alfred Binet, *Les Idées Modernes sur les Enfants*. Ernest Flammarion, Paris, 1909. p. 146.

The answer to this question may never be known for certain. In fact, the *IQ* itself is under attack and may in time be replaced by a more satisfactory measure. However, the vast majority of school children do not undergo such radical changes of environment during their school careers that the problem is of great practical significance to the teacher. Yet, as there are questions concerning motivation, emotional adjustment, optimum placement of pupils, and many others that bear significantly upon pupil performances not only on intelligence tests but also on achievement tests and in scholarship, the teacher should at least be aware of this controversial issue and some of its implications.

*Social class and the IQ.* Results from group intelligence tests have tended to show that children from certain socioeconomic groups attain higher mean scores than do children from other, and lower, socioeconomic groups. For example, rural children typically score lower than urban children, southern white pupils regularly score lower than northern white pupils, and children from working-class homes attain lower average scores than do those from homes at the professional and managerial levels.

Warner, Meeker, and Eells<sup>30</sup> showed that the cultural patterns of homes at different socioeconomic levels differ greatly. Davis, Havighurst, and others<sup>31</sup> obtained evidence to show that standard intelligence tests are not "culture free" but that they reflect the cultural biases of the upper-middle-class test constructors. Davis<sup>32</sup> indicated that differences of 8 to 12 *IQ* points for children from six to ten years of age and as high as 20 to 23 *IQ* points for children fourteen years of age between low and high socioeconomic groups reflect the cultural bias of the tests. He stated that culturally-fair tests used experimentally show pupils of low and high socioeconomic status to be closely similar in "innate intelligence" or "real intelligence."

These findings concerning the influence of culture, or home environment, on the *IQ* as obtained from standard intelligence tests appear not to be in disharmony with the evidence concerning the inconsistency of the *IQ*. If the results are borne out by more extensive

<sup>30</sup> W. Lloyd Warner, Marchia Meeker, and Kenneth Eells, *Social Class in America*. Science Research Associates, Chicago, 1949.

<sup>31</sup> *Ibid.* p. 26.

<sup>32</sup> Allison Davis, "Socio-Economic Influences on Learning." *Phi Delta Kappan*, 32:253-56; January 1951.



research, traditional methods of using the *IQ* in pupil guidance may well require revision.

*Future of the IQ.* It is apparent from the above discussion that the intelligence quotient is far from a perfect measure of brightness. It appears to be a more accurate measure for the years of middle childhood than for the first years of life or post-adolescent years. Its constancy seems to be somewhat in question. The influence of socioeconomic backgrounds may be significant. These weaknesses and others of a more technical nature raise logical questions concerning its continued and final acceptance as the best measure of brightness, although it is still one of the most satisfactory measures from which to predict success in school and is highly useful in pupil guidance. The alternative methods discussed below for indicating intelligence represent attempts to obtain a more satisfactory measure.

Freeman, after analyzing the problem carefully, stated that:

It may be true that the *IQ* is more convenient, but it is a question whether its inherent ambiguity does not make it better policy to adopt the statistically superior standard score and to educate teachers to understand and use it.<sup>33</sup>

Another attack on the *IQ*<sup>34</sup> recommended that the age-scale method of measuring intelligence be abolished, advocated the replacement of the mental age concept by a combination of measures from separate tests, and took the stand that the controversy concerning the constancy of the *IQ* is largely futile because its constancy or inconstancy does not depend upon fundamental issues but upon the manner in which tests provide means of obtaining the *IQ*.

Although the teacher should certainly understand the nature and proper uses of the *IQ*, he should also have some realization of its limitations, technical though they may be, and should be alert to the alternative methods for designating levels of intelligence which have been developed and which may be evolved in the future. The presentation of several alternatives below should take on additional significance in view of the apparent waning of prestige of the *IQ*.

<sup>33</sup> Frank N. Freeman, *Mental Tests: Their History, Principles and Applications*, Revised edition. Houghton Mifflin Co., Boston, 1939. p. 105.

<sup>34</sup> M. W. Richardson, "The Logic of Age Scales." *Educational and Psychological Measurement*, 1:25-34; January 1941.

## Personal constant (PC)

Heinis developed the personal constant<sup>35</sup> for the purpose of obtaining a measure that would be more accurate than the *IQ* for persons of very superior and very inferior intelligence levels. The measure, which he called the *per cent of average development*, but which is better known as the *personal constant*, is intended to give quantitative expression to the normal curve of mental growth in terms of growth units that have constant meaning at all age levels. The *PC* is computed by converting both the mental age and the chronological age to growth units by the use of a table of mental growth units,<sup>36</sup> dividing the *MA* value by the *CA* value, and multiplying by 100. Thus, the *PC* involves the substitution of growth units for *MA* and *CA* in the *IQ* formula.

Although Kuhlmann recommended that users of the *Kuhlmann-Anderson Intelligence Tests* employ it rather than the *IQ*,<sup>37</sup> and Hilden found that the *PC* fluctuates less than the *IQ*,<sup>38</sup> Cattell found the *IQ* to be definitely more constant for bright children and somewhat less constant for dull children than is the *PC*.<sup>39</sup> Freeman noted that the computation of the personal constant is more time-consuming than is that of the intelligence quotient, and indicated that the evidence now available concerning the values of the *PC* is inconclusive.<sup>40</sup>

## Index of brightness (IB)

The index of brightness is stated in the same form as the *IQ*. While its meaning is somewhat similar to that of the *IQ*, it is derived in quite a different manner. In this case, the pupil's relative brightness is expressed as a positive or negative deviation from the norm of pupils of his age. The difference between a pupil's score and the

<sup>35</sup> H. Heinis, "A Personal Constant." *Journal of Educational Psychology*, 17:163-86; March 1926.

<sup>36</sup> Arnold H. Hilden, *Table of Percent of Average Development Based on Mental Growth Units*. Educational Test Bureau, Minneapolis, 1936.

<sup>37</sup> F. Kuhlmann and Rose G. Anderson, *Instruction Manual: Kuhlmann-Anderson Intelligence Tests*, Fifth edition. Educational Test Bureau, Minneapolis, 1940. p. 17.

<sup>38</sup> Arnold H. Hilden, "A Comparative Study of the Intelligence Quotient and Heinis' Personal Constant." *Journal of Applied Psychology*, 17:355-75; August 1933.

<sup>39</sup> Psyche Cattell, "The Heinis Personal Constant as a Substitute for the *IQ*." *Journal of Educational Psychology*, 24:221-28; March 1933.

<sup>40</sup> Freeman, *op. cit.* p. 296.



norm for persons of the same chronological age is added to (if his score is above the norm) or subtracted from (if his score is below the norm) 100 to obtain his index of brightness. Otis, who used the measure for his *Quick-Scoring Group Tests of Mental Ability*, himself stated that the index of brightness has the same significance as an intelligence quotient.<sup>41</sup> Freeman, however, pointed out that the method by which the *IB* is derived makes improbable its consistency with the *IQ*.<sup>42</sup>

## Percentile scores

Percentile scores (also called centile scores) are frequently used to indicate a pupil's status in intelligence. This method is used particularly at the high-school and college levels, for the intelligence quotient, as has been pointed out above, is not as meaningful a measure for post-adolescent and adult years as it is for periods of childhood and adolescence. The percentile score describes a pupil's placement in an age or grade group in terms of the percentage of the group scoring lower than he does. The *American Council on Education Psychological Examinations* at both the high-school and college levels present norms for the interpretation of scores in terms of percentiles for different grade levels.

## Standard scores

Another type of measure that indicates a pupil's intelligence level in terms of his position within a certain age or grade group is based on the arithmetic mean and the standard deviation. Most frequently called standard scores, they have advantages over such relative measures of placement as percentile scores and are thought by some<sup>43</sup> to be superior to the *PC* as derived scores of intelligence. The *Merrill-Palmer Scale of Mental Tests* uses standard score norms. Terman and Merrill presented tables for the use of research workers and other persons in converting *IQs* obtained on the *New Revised Stanford-Binet Tests of Intelligence* into standard scores.<sup>44</sup>

<sup>41</sup> Arthur S. Otis, *Manual of Directions for Gamma Test: Otis Quick-Scoring Mental Ability Tests*. World Book Co., Yonkers, N. Y., 1937. p. 4.

<sup>42</sup> Freeman, *op. cit.* p. 300.

<sup>43</sup> Francis N. Maxfield, "Trends in Intelligence Testing." *Educational Research Bulletin*, 15:134-41; May 13, 1936.

<sup>44</sup> Terman and Merrill, *op. cit.* p. 42.

## 8 DISTRIBUTION OF INTELLIGENCE

It is important that the teacher know something of the manner in which intelligence is distributed if he is to make effective use of intelligence test results. The many reports of the distribution of intelligence show, however, that no single pattern of the distribution of intellectual ability can be expected to apply widely to different school situations. Typical groups of school children are not unselected, as might be supposed, but have been affected variously in their composition by many selective factors.

Intelligence can be conceived of both in terms of some such measure as the *IQ* and in terms of descriptions of the types of performance possible for persons of different intelligence levels. A distribution of intelligence quotients for an unselected group of children and the general descriptive terms used for different levels are presented here as an indication of the general distribution of intelligence.

TABLE 8. Distribution of intelligence quotients in a normal population <sup>45</sup>

Classification	IQ	Percentages of All Persons
Near genius or genius	140 and above	1
Very superior	130-139	2.5
Superior	120-129	8
Above average	110-119	16
Normal or average	90-109	45
Below average	80-89	16
Dull or borderline	70-79	8
Feeble-minded: moron	60-69	2.5
imbecile, idiot	59 and below	1

Table 8 shows the distribution of intelligence quotients for a normal population. Figure 20 presents the same data graphically. It will be noted that 45 per cent of the population fall within ten *IQ* points of the average *IQ* of 100. On the average, one person in each 100 is in the genius or near-genius class and one person in each

<sup>45</sup> Adapted from Terman and Merrill, *op. cit.*, p. 38-41. A standard deviation of 16.6 is used, in approximate accordance with results from Forms L and M of the *Stanford-Binet Examination* for all age groups between two and eighteen.



100 is in the very low feeble-minded group. About 10 to 12 per cent of the total may be considered as distinctly superior and 10 to 12 per cent as distinctly inferior. Persons at the highest level of feeble-mindedness, i.e., morons, are not uncommon in the lower grades of the school.

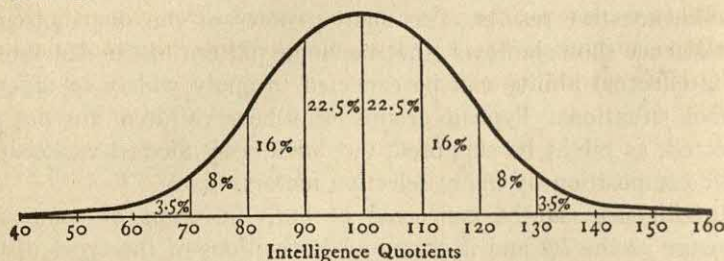


Fig. 20. Percentages of persons in a normal population at different levels of intelligence

## 9 DERIVED MEASURES RELATING INTELLIGENCE AND ACHIEVEMENT

It has been pointed out in the preceding chapters that intelligence tests are designed to measure primarily innate or inherited abilities and that achievement tests are intended to measure the results of education and experience. In one sense, then, intelligence tests can be considered as measuring *the capacity to learn* or *the potentialities for achievement* and achievement tests can be considered as measuring *what has been learned*. It seems natural, therefore, that an effort should be made to discover how well the individual is living up to his potentialities by comparing his performances on intelligence and achievement tests.

Two general procedures have been used for this purpose—those based on quotients and those based on differences. Two specific methods are discussed here. Only the first, which is discussed more fully, has come into wide use, but the other procedure is briefly presented so that the student may better grasp the problems involved in a reliable comparison of ability with achievement. As is pointed out later, measures of this type are highly questionable in their use with individual pupils, and even for use with pupil groups they must be interpreted with care and with regard for the many variables that condition their use if important pupil adjustments are to be made on the basis of the results.

### Accomplishment quotient (AQ)

The accomplishment quotient or achievement quotient, also sometimes called the accomplishment or achievement ratio, represents the relation between the educational level ( $EA$ ) and mental maturity ( $MA$ ) or between the relative educational development ( $EQ$ ) and relative brightness ( $IQ$ ) of a pupil. Therefore, the formula for the  $AQ$  is, in several adaptations,

$$AQ = 100 \frac{EA}{MA} = 100 \frac{EQ}{IQ} = 100 \frac{\frac{EA}{CA}}{\frac{MA}{CA}},$$

where  $EA$ ,  $MA$ , and  $CA$  indicate respectively the educational, mental, and chronological ages of the pupil expressed in months and  $EQ$  and  $IQ$  designate respectively his relative educational development and brightness.

For example, if a child has a mental age of ten years (120 months) and an educational age of nine years (108 months), his

$$AQ = 100 \frac{9.0}{10.0} = 100 \frac{108}{120} = 90.$$

If a pupil's achievement ( $EA$ ) is in keeping with his ability to learn ( $MA$ ), his  $AQ$  will be 100. Obviously, the indication of an  $AQ$  below 100 should be that the child is not working to capacity and an  $AQ$  of more than 100 should be impossible. However, a study of highly motivated instructional drives on certain content<sup>46</sup> showed that an  $AQ$  of more than 100 is attainable. It is certain, however, that no one can achieve at more than 100 per cent of his capacity. Therefore, it would appear that such accomplishment quotients result from norms on achievement tests that are not high in reliability.

There is evidence to show that higher accomplishment quotients are more frequently obtained in particular grade groups by the intellectually inferior than by the intellectually superior pupils.<sup>47</sup> This probably is true largely because of the fact that the instruc-

<sup>46</sup> W. E. Lessenger, *Motivation and the Accomplishment Quotient Technique*. University of Iowa Studies in Education, Vol. III, No. 2. University of Iowa, Iowa City, 1925.

<sup>47</sup> Harl R. Douglass and C. L. Huffaker, "Correlation between Intelligence and Accomplishment Quotient." *Journal of Applied Psychology*, 13:76-80; February 1929.



tional levels of most schools are geared to the average and inferior pupils and that the curriculum frequently does not have enough "top" adequately to interest and motivate superior pupils. Therefore, an  $AQ$  below 100 may indicate poor effort, a high  $IQ$ , or both, and an  $AQ$  of more than 100 may indicate unusual effort, a low  $IQ$ , or both.

Another weakness of the  $AQ$  is its low reliability,<sup>48</sup> which results from the fact that a ratio between two measures that are themselves not highly reliable for this comparison ( $EA$  and  $MA$ , or  $EQ$  and  $IQ$ ) cannot be highly reliable because the quotient of two unreliable measures is less reliable than either of the measures. In defense of these ages and quotients, it should be said that they have satisfactory degrees of accuracy for their normal uses but that they may not be sufficiently reliable for use in the ratios for obtaining the  $AQ$ .

A sound conclusion, growing out of the above and other more technical evaluations of the  $AQ$ , seems to be that its use with individual pupils is probably not justified but that it can satisfactorily be used for groups of pupils.

## Index of studiousness

An index of studiousness that attempts to relate ability to performance in the classroom has been proposed.<sup>49</sup> In its simplest form this measure is the difference between a pupil's rank in his class on intelligence and on achievement as they are measured by standardized tests. The index of studiousness is practically limited in comparable application to pupils within a class or instructional group and was recommended by its originator primarily for use in the high school.

## 10 GENERAL PROCEDURES FOR INTELLIGENCE TESTING

A large part of the continuing popularity of intelligence tests among teachers and supervisors may be traced to three main causes: (1) the tests themselves have been greatly improved in the accuracy

<sup>48</sup> J. Crosby Chapman, "The Unreliability of the Difference between Intelligence and Educational Ratings." *Journal of Educational Psychology*, 14:103-8; February 1923.

<sup>49</sup> Percival M. Symonds, *Measurement in Secondary Education*. Macmillan Co., New York, 1928. p. 521-25.

and analytical value of the resulting measures; (2) a larger proportion of school officers have become intimately acquainted with intelligence tests and testing procedures, with a correspondingly greater appreciation of the functions they serve; and (3) the changes in modern conceptions of education and attitudes toward it have made the utilization of such devices almost essential. Therefore, intelligence, aptitude, and group-factor tests are important tools of enlightened teaching procedure.

### **Administering and scoring intelligence tests**

During the early years of the intelligence testing movement the classroom teacher was given little part in the testing procedures and frequently was even denied access to the results. However, as teachers have become more conversant with intelligence testing techniques and the use of results, they have been given more responsibility in the administration and scoring of the tests and in the interpretation and use of results of group intelligence tests. The administration of individual intelligence tests and performance tests should remain a responsibility of the psychologist rather than that of the classroom teacher.

In many schools today, teachers administer, score, and quite often interpret the results of group intelligence tests. However, too great care cannot be taken by the teacher who participates in an intelligence testing program to understand the procedures for administering and scoring the tests and to follow the practices recommended by the test author, for it is only by such strict adherence to proper methods that reliability of the results is assured.

### **Care in the use of intelligence test results**

On the whole intelligence tests seem secure in the place they now hold as indispensable supporting tools for achievement tests, as valuable instruments for the more exact classification of pupils, and as guides to the teacher in matters of pupil behavior and conduct and to the pupil himself in certain vocational and related matters. There are, however, a few dangers attached to their careless or indiscriminate use which the teacher and administrator should guard against. The more important of these dangers are probably social in their character.



In the first place, there is the danger that may arise through giving publicity to the results of intelligence testing. In the long run, nothing but damage will ordinarily be done by using intelligence test results for any other than purely school purposes, and then they should be used in strict confidence. A second danger in the careless use of such tests lies in the effect that knowledge of his own intelligence may have on the individual. The safest practice is to restrict information concerning results of intelligence testing to responsible school officers and teachers in the main, to make such information available to parents only in occasional and well-considered instances where need arises, and to withhold such information from pupils themselves until they reach senior high school or perhaps even the college level. In no case does it seem justifiable to make intelligence quotients of individual pupils known to any persons other than their teachers and school officers, their parents, and themselves.

## 11 VALUES AND USES OF DIFFERENT TYPES OF TESTS

As tests of general intelligence, of aptitude and readiness, of group factors of intelligence, and of performance differ widely in type, mode of use, and nature of the resulting scores, it is inevitable that the situations in which they are most appropriately used must also differ.

### General intelligence tests

There is wide use for the results of general intelligence tests in the classroom. Results from group tests must be interpreted cautiously, however, for these indirect measures of adaptability or of ability to learn are often not highly reliable. Results that may safely be used for group interpretations may well be too unreliable for individual pupil interpretations. The best safeguard is to administer group tests frequently—perhaps every two or three years—during the school career of the pupil and to judge pupil intelligence more in terms of average intelligence quotients than in terms of the results from any one test, even the most recent, alone. For pupils who have very low or very high *IQs* and for pupils who are poorly adjusted to school, the administration of an individual intelligence test by a school psychologist is desirable. It is for the maladjusted, dull or borderline, and superior pupils that group test results are most likely

to be unreliable. Furthermore, it is for pupils of these types that the attainment of optimum adjustment in the school is the most difficult. Hence, the significance of results from individual intelligence tests in such cases is great.

*Individual diagnosis.* The intelligence test may prove especially valuable to the classroom teacher in assisting him to solve the problems relating to the unusual child. The pupil may be unusually bright, troublesome, dull, or in some other way quite out of the ordinary. The teacher may wish to know whether this child's typical responses reveal his real general ability, and whether or not the judgments of his former teachers and supervisors are correct. Intelligence tests will give information not obtainable in any other way.

The intelligence test, when given to an entire group, frequently uncovers a child of outstanding ability who has been content to go on with the group without revealing his real ability. Such tests invariably uncover cases of overlapping in ability just as achievement tests reveal cases of overlapping in school achievement. They may reveal children in the fourth and fifth grades with the mental ability of normal seventh- or eighth-grade pupils, or fourth- and fifth-grade pupils who in mental age are at the second- or third-grade level.

When children are discovered who are mentally far in advance of their place in school, readjustments of work should be made to match their abilities. This may be accomplished by (1) advancing them to a grade where their intelligence is given a real test, (2) placing them in rapidly moving classes, so that they may progress according to ability rather than by some fixed promotion scheme, or (3) declaring minimal requirements for the entire class in the units of work to be done and then expecting the brighter pupils to attack the various problems at higher levels and more intensively than could be expected of the class as a whole.

By proceeding along similar lines, the teacher will better understand the dull pupil because his difficulties can then be diagnosed more particularly and his strong points brought into relief. Quite frequently this results in an entire redirection and reorganization of his instruction. In any case, the intelligence test will assist both in explaining difficult cases and in revealing unsuspected general strengths and weaknesses.

*Educational guidance.* The use of intelligence test results for educational guidance is similar to but goes far beyond their use for individual pupil diagnosis. Pupils can be much more effectively



advised in their selection of courses and of curricula, and, by the same token, courses can better be adapted to their needs, if information is available concerning their intellectual levels. Pupils may be better qualified for certain types of courses or curricula than for others, in terms of their levels of intelligence. Evidence now available from some tests concerning ability levels in several areas or types of performance makes even more significant than formerly the possibilities of using intelligence test results in this manner.

*Vocational guidance.* The dividing line between educational and vocational guidance cannot be clearly drawn, for the first merges gradually into the second. Whereas educational guidance is of primary concern in the elementary school, even there it has its vocational implications. Vocational aspects of guidance assume an increasingly prominent position as the pupil progresses through junior and senior high school and in many instances nears the end of his school career. Although intelligence test results can be used with less confidence for vocational than for educational guidance, the information they furnish concerning the general intellectual abilities of pupils is of great value in vocational counseling.

*Class analysis and diagnosis.* Viewed from the standpoint of the teacher, achievement tests and intelligence tests are supplementary devices. After the teacher has given achievement tests and compared his class with the norms in a given subject, he is still in danger of making false assumptions about the significance of these results unless he has available further information such as is furnished by intelligence tests. He may credit himself with an excellent job of teaching when the innate brilliance of the group in his charge is such that, if they were given really adequate instruction, much superior results might have been achieved. Or it may be that the class falls so far below the norm that the teacher may feel that his teaching has proven a failure. This may also be an unwarranted assumption, for the class may be considerably below average in intelligence and cannot be expected to approximate the norms that are set up for a class of average ability. It is evident, then, that there is a need for some means of determining approximately the intellectual ability of the class. Intelligence tests meet this need. By giving one or more such tests, the teacher can determine with a fair degree of accuracy whether his class is up to the normal expectation in ability to master schoolwork.

## Specific intelligence tests

Much of what has been said concerning the uses of general intelligence tests applies also to aptitude and readiness tests. However, the specific nature of these types of tests limits their significance to certain uses that are correctly made of general intelligence tests.

Aptitude tests are valuable for individual pupil diagnosis, educational guidance, and vocational guidance. However, they have less significance for class analysis and diagnosis because they measure such specific abilities that individual pupil characteristics assume much greater importance than do characteristics of the class as a whole. Aptitude tests are primarily suited for use with pupils of the junior high school or higher levels, for the general and non-specialized type of course in the elementary school is less well-adapted to aptitude testing than are the more specialized courses of the high school and the college.

Readiness tests are useful for individual pupil diagnosis and for educational guidance but seem to have little significance for vocational guidance or class analysis and diagnosis. Such tests also have specific rather than general significance, so that the results from their use should be interpreted for the pupil as an individual rather than on the basis of the class group.

## Group-factor tests of intelligence

The group-factor tests, lying perhaps midway between general intelligence and aptitude tests in specificity, have uses similar to those outlined above. As these instruments have been developed comparatively recently, scores resulting from their use should be interpreted with caution and primarily by educational and vocational counselors until their validities for various purposes have become well established.

The bi-factor tests, most often supplying a verbal and a non-verbal score, doubtless distinguish two major factors of ability at any level from the intermediate grades to the college years. Boys regularly attain higher mean scores on the non-verbal sections than do girls, whereas the sex difference is typically in the opposite direction for verbal scores. Overlaps between the sexes are very great, however, so that many girls score far above the mean for boys



on the non-verbal tests and many boys surpass the mean for girls on the verbal tests by a wide margin. The diagnostic significance of these part scores seems sufficiently well established to warrant their cautious use in individual pupil diagnosis and for educational guidance when supporting evidence of other types is at hand. Their use in vocational guidance and for class diagnosis seems to be somewhat less appropriate.

Multi-factor tests of primary mental abilities and differential aptitudes have not yet resulted in a clear-cut and generally accepted list of ability factors, nor have the validities of the various part scores been established to the point where their predictive significance is well known. Somewhat more widely available for the high-school and college than for the intermediate grade levels, their major uses appear to be in the areas of individual diagnosis and vocational and educational guidance. It seems desirable at the present time to use results from multi-factor tests for pupil guidance only in conjunction with other data of well-established validity.

### Performance tests of intelligence

Performance tests are less frequently a tool of the classroom teacher than of the educational or vocational counselor. Pupils who have visual, language, or physical handicaps that preclude reliable testing of their abilities by group intelligence tests should be tested by individual intelligence scales or performance tests. The uses of results from performance tests do not differ significantly from the uses of group intelligence test results except that performance tests furnish less accurate measures of general intelligence than do group and individual intelligence tests and therefore should be employed with caution.

### Topics for Discussion

1. Give several of the most meaningful definitions of intelligence. Which definition is most acceptable to you? Why?
2. Briefly discuss and evaluate the three theories concerning the nature of intelligence that are presented in the chapter.
3. Show how a high score on an intelligence test affords a basis for inferring the existence of a high degree of mental ability.
4. What are the most appropriate uses for group intelligence tests? For individual intelligence tests?

5. Discuss the theoretical foundation upon which specific intelligence tests depend.
6. Distinguish between aptitude tests and readiness tests.
7. Upon what theoretical foundation do group-factor tests of intelligence rest?
8. Distinguish between bi-factor and multi-factor tests of intelligence.
9. For what purposes are performance tests ordinarily used?
10. Discuss fully the most commonly used measures of mental maturity and brightness.
11. To approximately what age does mental growth continue?
12. Does a person's intelligence quotient remain constant throughout his life? Give evidence to support your answer.
13. What would a culturally-fair intelligence test accomplish that standard intelligence tests may not accomplish?
14. Discuss the use of percentile scores and standard scores for the interpretation of intelligence test results.
15. How is intelligence distributed among the population as a whole?
16. What does the accomplishment quotient attempt to show? Discuss its defects and proper uses.
17. If a child has a *CA* of 12-6 and an *MA* of 15-10, what is his *IQ*? If his *EA* is 14-6, what is his *AQ*?
18. If the parents of a sixth-grade child who had a *CA* of 13-6 and an *MA* of 10-1 called upon you to discuss his poor work in arithmetic, what would you tell them about the causes of the child's deficiency in arithmetic?
19. List and discuss some of the ways in which intelligence test results are useful in the classroom.
20. Under what conditions, if any, do you think classroom teachers should be responsible for giving and scoring intelligence tests?
21. Propose a program to be followed in a school for the recording and use of intelligence quotients or other derived scores of intelligence.

## Selected References

- BINGHAM, WALTER V. *Aptitudes and Aptitude Testing*. New York: Harper and Brothers, 1937.
- BROOM, M. E. *Educational Measurements in the Elementary School*. New York: McGraw-Hill Book Co., Inc., 1939. Chapters 11-12.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 371-483.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p. 198-267, 428-65.



- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938. p. 95-114.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. p. 293-399, 627-750.
- CRONBACH, LEE J. *Essentials of Psychological Testing*. New York: Harper and Brothers, 1949. Chapters 6-10.
- FREEMAN, FRANK N. *Mental Tests: Their History, Principles and Applications*. Revised edition. Boston: Houghton Mifflin Co., 1939. Chapters 1-7, 9-11, 13-16.
- FREEMAN, FRANK S. *Theory and Practice of Psychological Testing*. New York: Henry Holt and Co., 1950. Chapters 3-8, 10, 16.
- FRIEDMAN, BERTHA S. "Intelligence Quotient." *Encyclopedia of Modern Education*. New York: Philosophical Library, 1943. p. 407-10.
- FROELICH, CLIFFORD P., AND DARLEY, JOHN G. *Studying Students: Guidance Methods of Individual Analysis*. Chicago: Science Research Associates, 1952. Chapter 10.
- GOODENOUGH, FLORENCE L. *The Measurement of Intelligence by Drawings*. Yonkers, N. Y.: World Book Co., 1926.
- GOODENOUGH, FLORENCE L. *Mental Testing: Its History, Principles, and Applications*. New York: Rinehart and Co., Inc., 1949. Chapters 21-23.
- GREENE, EDWARD B. *Measurements of Human Behavior*. Revised edition. New York: Odyssey Press, 1952. Chapters 5-6, 8-11.
- HOLZINGER, KARL J. "Factor Analysis." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 429-33.
- HULL, CLARK L. *Aptitude Testing*. Yonkers, N. Y.: World Book Co., 1928.
- HUMPHREYS, LLOYD G., AND BOYNTON, PAUL L. "Intelligence and Intelligence Tests." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 600-12.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. Chapters 14-15.
- MCMEMAR, QUINN. *The Revision of the Stanford-Binet Scale: An Analysis of Standardization Data*. Boston: Houghton Mifflin Co., 1942.
- MURSELL, JAMES L. *Psychological Testing*. Second edition. New York: Longmans, Green and Co., 1949. Chapters 3-7, 9-10.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. Chapter 11.
- ORLEANS, JACOB S. *Measurement in Education*. New York: Thomas Nelson and Sons, 1937. Chapter 3.

- REMMERS, H. H., AND GAGE, N. L. *Educational Measurement and Evaluation*. New York: Harper and Brothers, 1943. Chapters 14-15.
- STODDARD, GEORGE D. *The Meaning of Intelligence*. New York: Macmillan Co., 1943. Chapters 1, 11-12, 15-16.
- STODDARD, GEORGE D., chairman. *Intelligence: Its Nature and Nurture*. Thirty-Ninth Yearbook of the National Society for the Study of Education, Parts I and II. Bloomington, Ill.: Public School Publishing Co., 1940.
- SUPER, DONALD E. *Appraising Vocational Fitness by Means of Psychological Tests*. New York: Harper and Brothers, 1949. Chapters 4, 6, 8-11, 15.
- TERMAN, LEWIS M., AND MERRILL, MAUD A. *Measuring Intelligence: A Guide to Administration of the New Revised Stanford-Binet Tests of Intelligence*. Boston: Houghton Mifflin Co., 1937.
- THURSTONE, L. L. *Primary Mental Abilities*. Psychometric Monograph Series, No. 1. Chicago: University of Chicago Press, 1938.
- TRAXLER, ARTHUR E. *Techniques of Guidance: Tests, Records, and Counseling in a Guidance Program*. New York: Harper and Brothers, 1945. Chapter 4.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapters 5-7.
- WOOD, BEN D., AND HAEFNER, RALPH. *Measuring and Guiding Individual Growth*. New York: Silver Burdett Co., 1948. p. 263-74, 311-20.



## ***Using Personality Instruments and Techniques***

THE FOLLOWING ASPECTS of personality and its measurement are discussed in this chapter:

- A. Nature of personality.
- B. Methods of personality measurement.
- C. Nature and measurement of attitudes.
- D. Nature and measurement of interests.
- E. Significance and measurement of emotional adjustment.
- F. Measurement of total personality.

Teachers are expected to understand their pupils, and through this understanding to increase the efficiency of their teaching. To the teacher of some decades ago, all pupils were essentially alike. The modern teacher should have a knowledge of child psychology and the nature of individual differences in intelligence, achievement, and other important aspects of behavior. Many teachers give too little attention to the personality aspects of child behavior, preferring to work with the more readily observable and more tangible phases of behavior such as those treated in the chapters on achievement and intelligence testing. It is probable that teacher-education institutions even today too infrequently provide teachers in training with adequate instruction concerning pupil personality in a functional sense. Wherever the fault may lie, attention is increasingly being directed toward the more effective adjustment of the school to the needs of the child and of the child to life. Thus efficient teaching demands

more than a chance and casual acquaintance with personality testing techniques.

## 1 NATURE OF PERSONALITY

Man has for centuries been aware of differences among individuals and has made many attempts to classify them. As early as 2000 B.C., Theophrastus divided men into thirty universal types,<sup>1</sup> of which the dissimulator, the flatterer, the chatterer, and the rustic are representative. Hippocrates, several centuries B.C., distinguished persons of the sanguine, choleric, melancholic, and phlegmatic characteristics and explained these various types of temperaments by excesses of the bodily fluids or "humors" he called blood, yellow bile, black bile, and phlegm respectively.<sup>2</sup> Palmistry, phrenology, numerology, and graphology have long made claims concerning their ability to diagnose personality. More recently, Kretschmer divided men by their physical characteristics into four types distinguishable by certain general personality characteristics,<sup>3</sup> and Berman emphasized the influence of secretions from the endocrine or ductless glands upon personality.<sup>4</sup> Garrett referred to physiognomy, body type theories, and glandular theories as impressionistic and noted the contrast between their extravagant claims and the exceedingly meager results they produce.<sup>5</sup>

Jung distinguished extrovertive and introvertive types of individuals,<sup>6</sup> and his classification has to a considerable degree found its way into popular usage. More recently still, however, psychologists have increasingly turned their attention to the study of and attempts to measure personality. The concept of types evidenced in most of the early and many of the rather recent attempts to evaluate personality has largely been abandoned by modern per-

<sup>1</sup> Richard Aldington, editor, *A Book of Characters from Theophrastus*. E. P. Dutton and Co., New York, 1924.

<sup>2</sup> Laurance F. Shaffer, *The Psychology of Adjustment*. Houghton Mifflin Co., Boston, 1936. p. 284.

<sup>3</sup> Ernst Kretschmer, *Physique and Character*. Harcourt, Brace and Co., New York, 1925.

<sup>4</sup> Louis Berman, *The Glands Regulating Personality*. Macmillan Co., New York, 1921.

<sup>5</sup> Henry E. Garrett, *Great Experiments in Psychology*, Third edition. Appleton-Century-Crofts, Inc., New York, 1951. p. 175-82.

<sup>6</sup> C. G. Jung, *Psychological Types*. (Translated by H. G. Baynes.) Harcourt, Brace and Co., New York, 1923.



sonality testers, for personality types are inconsistent with the "normal curve" distribution that has been found to apply to personality traits as well as to intelligence and achievement.

Personality was at one time thought to be largely if not entirely the result of biological inheritance. However, most authorities today prefer the view that it is the resultant of both hereditary and environmental factors.<sup>7</sup> Psychoanalysts believe that many of the personality difficulties found among adults are caused primarily by experiences of early childhood, in many cases forgotten by the adult. If personality characteristics are the result in significant measure of the environment, which seems a justifiable conclusion, it is important for the teacher to be alert to the influence of the school in shaping the personality of the child as well as to its potentialities for correcting the maladjustments that pupils may have acquired prior to school entrance.

### Definitions of personality

Personality is the most inclusive term that can be used in the discussion of human behavior. Psychologists are not in complete agreement concerning the meaning of the term, but they recognize that personality describes more fundamental types of human behavior than the surface evidences by which the man on the street evaluates it. In general psychological definitions of personality explain what personality is in terms of the types of human behavior thought to contribute to it. Psychologists agree roughly upon these components of personality, but they usually resort to indirect methods of defining the term.

Shaffer stated that the "personality traits of an individual are his persistent habits toward making certain types of adjustments rather than other kinds."<sup>8</sup> Traxler considered the term to include the "sum total of an individual's behavior in social situations."<sup>9</sup>

These statements concerning personality seem to describe as well as possible in a non-technical manner what personality is. They perhaps represent the most meaningful view of personality for

<sup>7</sup> Willard C. Olson, "Personality." *Encyclopedia of Educational Research*, Revised edition. Macmillan Co., New York, 1950. p. 807.

<sup>8</sup> Shaffer, *op. cit.* p. 132.

<sup>9</sup> Arthur E. Traxler, *Techniques of Guidance: Tests, Records, and Counseling in a Guidance Program*. Harper and Brothers, New York, 1945. p. 100.

teachers and other persons who are not technical workers in the field of personality study. It should be kept clearly in mind that the behavior of the individual is controlled by his personality and at the same time furnishes the evidence by which his personality can in part be evaluated.

### Aspects of personality

If personality is most satisfactorily described at present in terms of how it is constituted, it is understandable that approaches to personality study and measurement have been largely in terms of personality traits. Psychologists divide personality into many areas for study. However, the aspects of personality useful to the teacher can well be listed under fewer headings, although any classification must be largely arbitrary. The phases of personality treated in this chapter are grouped under the headings of attitudes, interests, emotional adjustment, and total personality. It is believed that these are the areas of greatest present significance to the teacher.

Although the preceding statements include intellectual and physical traits as components of personality, these traits are not generally considered when personality measurement is undertaken. They are measured by different techniques in established areas of testing. Consequently, although the psychology of personality rightly deals with their findings and no one should lose sight of the contributions of intellectual and physical traits to an individual's development, these areas are not of direct concern here.

## 2 TECHNIQUES OF PERSONALITY MEASUREMENT

Personality is measured by several different types of approaches. Among those most commonly used are (1) free association, (2) direct observation of behavior, (3) rating scales, and (4) personal reports. Although all of these methods can be used by an intelligent classroom teacher, it is probable that observation of behavior and personal reports are the methods most practicable and useful in the typical classroom. Each of these methods is discussed briefly in this section of the chapter. In the later sections various methods of measurement are discussed in terms of their uses for the evaluation of attitudes, interests, emotional adjustment, and total personality.



Personality testing is probably the newest area of measurement that bears directly upon the work of the classroom teacher. Although achievement and, to a less extent, intelligence, are subject to quantitative measurement and now have rather widely accepted technical terminologies to aid in the interpretation of testing results, such is not the case for personality testing. Results from personality tests are often difficult to interpret because the area has practically no derived scores, such as the educational age and the intelligence quotient, which have commonly accepted meanings. The effect of this situation is that personality test results must be interpreted largely in terms of the special types of derived scores and norms provided for the particular test or technique used and then frequently by qualitative rather than quantitative statements.

### Association methods

An association method was one of the earliest to be employed in the measurement of behavior, for it apparently was first used by Galton as early as 1879.<sup>10</sup> Its development has occurred mainly since 1910 in the modern sense, however.

Two association methods are now being quite widely used in the study of personality: (1) verbal association techniques, and (2) visual stimulus techniques. Although word association methods were known long before modern projective methods evolved, both types are now known as projective techniques.

Verbal associations are established when the person to whom a word is spoken responds with the first word that enters his mind. Other free-association procedures are based on completions given to incomplete sentences and to partly told stories.

The best-known visual stimulus methods are based on free responses to inkblots and pictures. In these situations the subject is to respond by telling what he is reminded of or what he sees in each. Both the nature of the responses and the manner in which they are given furnish considerable evidence to the experienced psychologist on which to base inferences concerning emotional disturbances in the subject.

<sup>10</sup> Francis Galton, "Psychometric Experiments." *Brain*, 2:149-62; July 1879.

## Observational methods

Several different methods based on the observation of individual pupil behavior have been suggested and successfully applied. They all probably require an ability that few teachers have but most can acquire. Untrained teachers make use of their own interpretations of events they observe, whereas objectivity is attained only by a rather rigid account of what actually occurred. The characteristics of objective observational methods highly useful in the study of pupil personality and adjustment make it inadvisable for inexperienced teachers to attempt to make more than experimental use of them until some experience in observation has been acquired.

The two common observational procedures most applicable in the school are (1) directed observation and (2) the anecdotal method. The first, because the observation is directed toward a particular pupil or pupil group under specified conditions, is a laboratory rather than a classroom procedure. The second, however, uses the results from observations of pupil behavior made at any time, and therefore is definitely a classroom method of evaluation.

Certain observational procedures are generally known as projective methods. In using these, the child is presented with such materials as sand, clay, toys, or paints and his use of the material presented is carefully observed by the psychologist. Much is revealed to the experienced observer concerning the conscious and even unconscious motives, attitudes, interests, and needs of the individual by this approach.

## Rating scales

Rating scales are widely used in the evaluation of pupil personality. In this procedure, the teacher or some other person intimately acquainted with the pupils rates them on personality traits in terms of the manner in which the individuals have impressed the rater. Obviously, the judges should know intimately the pupils they are rating. Most rating methods suffer in accuracy because some raters tend to be too lenient whereas others are too critical. They are less accurate for use with intangible traits, upon which observers usually vary rather widely in their evaluations, than in such readily observable characteristics as neatness and cleanliness.

Widely used among rating techniques are the graphic rating scale



and variations of that form. In this procedure, the judge places a check mark at a certain position on a line to indicate his evaluation of the person he is rating. The line may be divided into five (or some other number of) sections designated superior, good, average, poor, and inferior, or meaning may be given to positions on the scale by other and more definitely descriptive terms. Again, there may be designations at occasional intervals beneath the line to indicate specifically for each trait varying evidences of its possession by the person being rated. Several personality rating scales are discussed later in this chapter.

### Personal reports

The personal report method makes use of what are variously called scales, inventories, questionnaires, and blanks. The responses are given, or the instruments are filled out, by the pupils themselves. As many of the items on these instruments request highly personal responses, the personal report method of measuring personality suffers from the fact that pupils sometimes reply as they think they should reply rather than as they truly react to the various items. Most persons are hesitant in revealing their inner personalities to other persons freely. In fact, the customs of civilized society place something of a premium upon the ability to hide or disguise emotions, likes and dislikes, attitudes, and other reactions in many situations. Therefore, it is not surprising that pupils sometimes fail to answer personality inventory items truthfully. Despite this major weakness, personal report instruments for the measurement of personality are of considerable value in their classroom uses.

## 3 MEASUREMENT OF ATTITUDES

A significant portion of the teacher's time in the classroom is directly or indirectly devoted to the development in pupils of desirable social attitudes and modes of behavior. Illustrations are the emphasis in the school upon good citizenship, cooperation with others, intellectual honesty, and the scientific attitude. Furthermore, many courses in the school attempt to develop attitudes that are in many cases more specific than those mentioned above. For example, the teacher of English attempts to develop favorable attitudes toward correct usage and good literature in his pupils, and the teacher

of civics to develop democratic ideals and belief in democratic institutions. Lists of course objectives include many such attitudes, ideals, or beliefs which the school strives to develop or improve in the pupil. It seems important, then, that teachers be conversant with instruments for measuring attitudes of various types.

## The nature of attitudes

Thurstone and Chave defined an attitude as "the sum total of a man's inclinations and feelings, prejudice or bias, preconceived notions, ideas, fears, threats and convictions about any specific topic."<sup>11</sup> An attitude is a state of readiness that exerts a directive, and sometimes a compulsive, influence upon an individual's behavior.

Attitudes may be either general or specific. For example, a person who has a general attitude of liberalism may behave in a highly conservative manner in a particular situation in which his personal welfare may be threatened. An attitude of conservatism is general, but an attitude toward a certain person is specific. This brief indication of the nature of attitudes will furnish the student sufficient background concerning the psychological characteristics of attitudes for the brief consideration of measuring instruments presented in this chapter.

## Methods of attitudes measurement

Attitudes are measured by several different methods, among the most common of which are the scale or questionnaire and the interview. As the teacher ordinarily has more use for the attitudes scale than for the interview, only brief mention will be made of the interview as a device for determining attitudes and opinions.

Attitudes scales. The two series of attitudes scales most widely used and known are the *Thurstone Scales for the Measurement of Social Attitudes* and the *Generalized Attitudes Scales* devised by Remmers and his associates. The former measure attitudes toward areas where differences of opinion exist, such as censorship, immigration, unions, war, capital punishment, and the movies, whereas the latter measure such attitudes as those toward any social institution, any racial or national group, any vocation, and any school subject.

<sup>11</sup> L. L. Thurstone and E. J. Chave, *The Measurement of Attitude*. University of Chicago Press, Chicago, 1929. p. 6-7.



Sample items from the *Around the World* attitudes inventory for pupils from the sixth to the tenth grade are presented in an accompanying illustration. This inventory measures attitudes toward various phases of international relations, war, patriotism, and agencies for peace.

### Sample items from "Around the World" Attitudes Inventory <sup>12</sup>

#### I. YES OR NO?

Below are a number of statements, some of which are true and some are false. Underline YES if you think a statement is true. If you think it is not true, underline NO.

1. Most foreigners are less intelligent than Americans . . . YES NO
7. In most American homes there are things made out of material that came from other countries . . . . . YES NO
10. In modern warfare people who live far from the fighting are often in great danger . . . . . YES NO

Another illustration of an attitudes measurement technique is that of the *Health Attitudes Inventory*. This is one of the six health inventories developed by the Cooperative Study in General Education.

### Excerpt from Health Attitudes Inventory <sup>13</sup>

**Directions:** Read each of the following statements and on the appropriate line of the answer sheet blacken the space under

- A if you agree with the whole statement,  
 D if you disagree with the whole statement,  
 U if you are uncertain how you feel about the whole statement.

#### STATEMENTS

1. People lose their hair because they do not take care of it in a way which is known to be right.
2. Before shaking hands with a person, one should if possible make sure that he does not have any skin disorders.
3. If two people maintain that the use of certain foods cause different effects on their skins, one or the other or both probably made incorrect observations.
4. Ordinarily there is little danger of catching a disease by wearing new clothes before they are laundered.

<sup>12</sup> Adelaide T. Case and Paul M. Limbert, *Around the World*. Published by Association Press, 1932.

<sup>13</sup> Cooperative Study in General Education, *Health Attitudes, Health Inventory* No. IV. Published by Educational Testing Service, 1950.

*The interview.* The method of determining attitudes by the use of the interview is similar to the method of general interviewing discussed in Chapter 9. However, the interview for purposes of attitudes measurement is usually restricted to rather direct and somewhat standardized questioning on the particular issue toward which attitudes are being measured.

## 4 MEASUREMENT OF INTERESTS

The attention of classroom teachers has increasingly turned of late years to pupil interests, as a result of the emphasis now placed upon the adaptation of the school offerings to the abilities, needs, and interests of pupils. Furthermore, the vocational and avocational interests of children have increasingly received attention during the last few decades as a means of aiding the pupil in the selection of his school courses and curricula and his life vocation. The school obviously cannot adapt its offerings to pupil interests and guide pupils in their selection of courses in terms of interests if the nature of those interests is unknown.

### The nature of interests

Interests today are most often classified in terms of the objects and activities from which the individual obtains satisfaction.<sup>14</sup> Thus, a person is interested in football but cares very little for tennis, or he is interested in music but is not interested in the drama. It is in this non-technical manner of considering interests that measurement in this field can be most meaningful for the teacher.

### Methods of interests measurement

Interests are subject to measurement both by standardized interests inventories and by informal methods. Brief mention will be made here of informal testing methods. Interests inventories will be treated somewhat more fully.

*Interests inventories.* It is perhaps because of the fact that inventories of pupil interests in objects and activities are so easy to make that standardized inventories most often deal with interests from

<sup>14</sup> Douglas Fryer, *The Measurement of Interests*. Henry Holt and Co., New York, 1931. p. 15.



the standpoint of their predictive or diagnostic values for important types of behavior. The result is that certain types of instruments perhaps best classified as inventories do not differ greatly from attitudes scales and that other types are very similar to, or in effect may be, adjustment inventories.

The Pressey *Interest-Attitude Test*, a brief excerpt from which appears on page 60, is an illustration of the first type. Another illustration in a non-vocational interest area is the *Health Interests Inventory*, from which an excerpt appears in an accompanying illustration. A comparison of this and the excerpt from the *Health Attitudes Inventory* on page 286 may help to clarify the minor distinction between attitudes scales and general interests inventories.

### Excerpt from Health Interests Inventory <sup>15</sup>

**Directions:** Read each of the questions below and on the appropriate line of the answer sheet blacken the space under

- A if the question interests you and you feel it should be dealt with in school,
- B if the question interests you but you feel it should not be dealt with in school,
- C if the question does not interest you.

1. Do certain diseases of the skin result from beauty-parlor or barbershop treatments?
2. Are pimples caused by poor digestion?
3. If the skin is dry and becomes itchy after bathing, what should be done?
4. What is the proper treatment for boils?
5. How can athlete's foot be cured?

Probably the best-known measuring instrument in the field is the *Strong Vocational Interest Blank*. This inventory is not intended for use below the senior high school and college levels because of the transient nature of interests in vocations at lower age levels. In common with many other interests inventories, the Strong blank has separate forms for young men and young women in order to provide for the types of sex differences usually found to exist in interests.

Persons taking the Strong blank are asked to respond to items dealing with the following: (1) occupations, (2) school subjects, (3) amusements, (4) activities, (5) peculiarities of people, (6) order of preference of activities, (7) comparison of interest between two

<sup>15</sup> Cooperative Study in General Education, *Health Interests, Health Inventory* No. III. Published by Educational Testing Service, 1950.

items, and (8) rating of present abilities and characteristics. They designate their interests on a three-point scale for most of the items to indicate their degree of liking. The samples of the accompanying illustration show the nature of differences in the men's and women's occupations items, the identical nature of items on present abilities and characteristics in the men's and women's forms, and two of the methods of responding to the items.

Scoring of the instrument requires quite lengthy procedures involving the use of varying positive, zero, and negative weights differing for the same response according to the particular vocation for which the blank is being scored. The men's form can be scored for 35 different occupations, as well as for several occupational groups and two special occupational indices.<sup>16</sup> The women's form can be scored for 16 occupations and one special occupational index.<sup>17</sup>

*Informal measurement of interests.* Educational literature of recent years includes many reports of interests studies in a variety of school subjects and areas of behavior. Among the fields for which such studies have appeared in considerable number are reading interests in books, magazines, and newspapers; play interests; interests in the movies, television, and radio; and interests in various subjects of the elementary school and high school. Reference to such sources will furnish the teacher much information concerning interests of various pupil groups.

However, the teacher can obtain direct information concerning the interests of his pupils by informal methods. Questioning individual pupils and class groups about their interests is a simple procedure, and one productive of considerable information. The teacher may, however, have the pupils write about their interests or list them without discussion. Again, he may prepare and distribute to the pupils a list of books, of magazines, of recreational activities, or of any one of a number of other types of objects and activities and then ask the pupils to check those in which they are interested. In any of these procedures, it is wise to limit the investigation of interests to one area rather than to attempt a complete inventory of pupil interests at one time.

<sup>16</sup> Edward K. Strong, Jr., *Manual for Vocational Interest Blank for Men*. Stanford University Press, Stanford University, Cal., August 1938.

<sup>17</sup> Edward K. Strong, Jr., *Manual for Vocational Interest Blank for Women*. Stanford University Press, Stanford University, Cal., October 1938.



Excerpts from Strong Vocational Interest Blanks<sup>18</sup>

**Occupations.** Indicate after each occupation listed below whether you would like that kind of work or not. Disregard considerations of salary, social standing, future advancement, etc. Consider only whether or not you would like to do what is involved in the occupation. You are not asked if you would take up the occupation permanently, but merely whether or not you would enjoy that kind of work, regardless of any necessary skills, abilities, or training which you may or may not possess.

Draw a circle around L if you like that kind of work

Draw a circle around I if you are indifferent to that kind of work

Draw a circle around D if you dislike that kind of work

Work rapidly. Your first impressions are desired here. Answer all the items. Many of the seemingly trivial and irrelevant items are very useful in diagnosing your real attitude.

1 Actor (not movie) .....	L	I	D	1 Actress (movie) .....	L	I	D
2 Advertiser .....	L	I	D	2 Actress (stage) .....	L	I	D
3 Architect .....	L	I	D	3 Accountant .....	L	I	D
4 Army Officer .....	L	I	D	4 Advertiser .....	L	I	D
5 Artist .....	L	I	D	5 Architect .....	L	I	D
6 Astronomer .....	L	I	D	6 Artist .....	L	I	D
7 Athletic Director .....	L	I	D	7 Artist's Model .....	L	I	D
8 Auctioneer .....	L	I	D	8 Athletic Director .....	L	I	D

**Rating of Present Abilities and Characteristics.** Indicate below what kind of a person you are right now and what you have done. Check in the first column ("Yes") if the item really describes you, in the third column ("No") if the item does not describe you, and in the second column (?) if you are not sure. (Be frank in pointing out your weak points, for selection of a vocation must be made in terms of them as well as your strong points.)

	YES	?	NO
342 Usually start activities of my group .....	( )	( )	( )
343 Usually drive myself steadily (do not work by fits and starts) .....	( )	( )	( )
344 Win friends easily .....	( )	( )	( )
345 Usually get other people to do what I want done .....	( )	( )	( )

<sup>18</sup> Edward K. Strong, Jr., (1) *Vocational Interest Blank for Men*, Revised, and (2) *Vocational Interest Blank for Women*. Published by Stanford University Press, 1938 and 1933.

## 5 MEASUREMENT OF EMOTIONAL ADJUSTMENT

Every individual faces the problem of adjusting himself to a none-too-benign environment. Persons who are successful in adapting themselves to their environments are well adjusted; those who fail in this adaptation become maladjusted. The school seeks to improve the adjustment of its pupils by furnishing them important learning opportunities and experiences. However, it must go beyond learning in the classroom sense and attempt to bring about the best possible form of adjustment between the individual and his environment in terms of his total personality.

The measurement of adjustment is an extremely comprehensive task. In its broad sense such measurement implies the use of all types of devices that will furnish information concerning the child and his backgrounds of heredity and environment. The discussion of adjustment in this section applies primarily to emotional adjustment. Although this is a fundamentally important issue, because of the fact that maladjustment seems to have consequences of great importance in the emotional life of the individual, the measurement of emotional maladjustment should not be regarded as the sole approach to this problem. The discussion in certain portions of Chapter 9 deals with adjustment in a somewhat broader sense than the treatment given in this section.

### Causes and symptoms of maladjustment

Maladjustment may arise when an individual is frustrated in the satisfaction of his fundamentally important aims, motives, or goals. It is the result of a lack of balance between the difficulties the individual encounters in his environment and his ability to meet the difficulties successfully. The underlying causes may be of many types, and frequently they are very elusive. Frustration itself is a result, not a cause. The effects, or results, are much more readily determined than are the causes. Symptoms of maladjustment may fairly readily be observed by the teacher who has insight into pupil behavior, but the determination of causes underlying maladjustment is often a task for the clinical psychologist. Although some alleviation of maladjustment may be accomplished without knowledge of its causes, effective remediation depends upon a knowledge of and ability to cope successfully with the true causal factors.



## Methods of adjustment measurement

The importance of an awareness by the teacher of existent emotional maladjustments in his pupils should be apparent from the preceding discussion. Such recognition of maladjustments should be accompanied by evidence concerning their nature, and, if possible, their causes. Adjustment inventories serve the first two purposes of pointing out the existence of and nature of existing maladjustments quite adequately in many instances, but they probably do not accomplish the third purpose, of discovering the causes of maladjustments. They frequently, however, furnish evidence that will greatly facilitate further study of maladjusted pupils in the attempt to determine causes and then to eliminate them.

Three general procedures are probably most often used in the measurement of adjustment—personal report blanks, rating scales, and projective techniques. Each of these methods is discussed briefly and illustrated by a few representative instruments in the following pages.

### From Rogers Test of Personality Adjustment <sup>19</sup>

Suppose that just by wishing you could change yourself into any sort of person. Which of these people would you wish to be? Write a "1" in front of your first choice, a "2" in front of your second choice, and a "3" in front of your third choice:

- |                            |                      |
|----------------------------|----------------------|
| (a) _____ a housewife      | (n) _____ a fireman  |
| (b) _____ a teacher        | (o) _____ a poet     |
| (h) _____ a business woman | (t) _____ an actress |
| (l) _____ an aviator       | (y) _____ a salesman |
| (m) _____ a captain        | (z) _____ an artist  |

Is there any other sort of person you would like to be? If there is, write it here: \_\_\_\_\_

*Personal report blanks.* By far the majority of adjustment inventories make use of the personal report method, by which pupils are asked to give answers to a variety of questions. The considerable quantity of adjustment inventories and the wide variety of response methods they use precludes any more comprehensive treatment here than brief descriptions and illustrations of a few of them.

<sup>19</sup> Carl R. Rogers, *A Test of Personality Adjustment for Girls*. Published by Association Press, 1931.

An illustration from the *Bell Adjustment Inventory*, an instrument for measuring (1) home, (2) health, (3) social, and (4) emotional adjustment, was given on page 60. It will not be discussed further here. Sample items from the girls' form of the *Rogers Test of Personality Adjustment* are given herewith to illustrate procedures used in measuring the adjustment of elementary-school children. This inventory, for use with girls from nine to thirteen years old, is devised to measure adjustment of the girl toward other children, toward her family, and toward herself. The comparable form for boys is not illustrated here.

The *Aspects of Personality* inventory measures the temperament and personality traits of children in Grades 4 to 9 by the use of items of the type shown in the accompanying illustration. The inventory yields scores that can be translated into percentiles on an ascendance-submission, an extroversion-introversion, and an emotionality scale.

Excerpt from *Aspects of Personality* <sup>20</sup>

SECTION III		III	
1. I like to go to the movies.....	<input type="checkbox"/> S <input type="checkbox"/> D	1	
2. I think most children like to make fun of me.....	<input type="checkbox"/> S <input type="checkbox"/> D	2	
3. I get angry about nothing.....	<input type="checkbox"/> S <input type="checkbox"/> D	3	
4. I get so angry I can't talk.....	<input type="checkbox"/> S <input type="checkbox"/> D	4	

The *Guess Who Test*, illustrated by the sample given below, is intended for use in measuring a child's reputation among his fellows. The test, for use from Grade 5 to Grade 8, requests pupils to list their classmates who particularly fit the brief portraits presented to them. It is possible to obtain a total reputation score for each pupil in a class from the results.

Excerpt from *Guess Who Test* <sup>21</sup>

Here are some little word-pictures of children you may know. Read each statement carefully and see if you can guess who it is about. It might be about yourself. There may be more than one picture for the same person. Several boys and girls may fit one picture. Read each statement. Think over your classmates and write after each statement

<sup>20</sup> Rudolf Pintner and others, *Aspects of Personality*. Published by World Book Co., 1937.

<sup>21</sup> *Guess Who Test*. Published by Association Press, 1930.



the names of any boys or girls who may fit it. If the picture does not seem to fit anyone in your class, put down no names but go on to the next statement. Work carefully and use your judgment.

1. Here is the class athlete. He (or she) can play baseball, basketball, tennis, can swim as well as any, and is a good sport.
- 
- 

The *Mooney Problem Check List*, from which an excerpt is shown in an accompanying illustration, differs from many personal report forms in that it makes no provision for formal pupil scores and no norms are provided. Since its major uses are in counseling, surveying pupil problems, and research, the use of indicated problems, simple counts of problems by areas, and summaries of problems for groups of pupils constitute the recommended bases for interpretation of results. Local norms are considered to be of greater significance than national norms. Therefore, it is suggested by the publisher that they be derived as desired.

### Excerpt from Mooney Problem Check List <sup>22</sup>

**DIRECTIONS:** Read the list slowly, and as you come to a problem which troubles you, draw a line under it.

- 
- |                                   |                                      |
|-----------------------------------|--------------------------------------|
| 1. Often have headaches           | 36. Too short for my age             |
| 2. Don't get enough sleep         | 37. Too tall for my age              |
| 3. Have trouble with my teeth     | 38. Having poor posture              |
| 4. Not as healthy as I should be  | 39. Poor complexion or skin trouble  |
| 5. Not getting outdoors enough    | 40. Not good looking                 |
| 6. Getting low grades in school   | 41. Afraid of failing in school work |
| 7. Afraid of tests                | 42. Trouble with arithmetic          |
| 8. Being a grade behind in school | 43. Trouble with spelling or grammar |
| 9. Don't like to study            | 44. Slow in reading                  |
| 10. Not interested in books       | 45. Trouble with writing             |

*Rating scales.* Two rating scales that are of major use in locating maladjusted pupils are briefly commented upon and illustrated here. Although these scales have the same general purposes as the personal report blanks discussed above, the two types of adjustment measures differ greatly in method.

The *Haggerty-Olson-Wickman Behavior Rating Schedules* are illustrated by the few following items. Although this scale is similar

<sup>22</sup> Ross L. Mooney, *Mooney Problem Check List*, Junior High School Form, 1950 revision. Copyright by Psychological Corporation, 1950.

in general appearance to a graphic rating scale, it differs in that the two extremes do not necessarily represent the most and the least desirable situations. Instead, the numbers 1 to 5, variously spaced

Excerpt from Haggerty-Olson-Wickman Behavior Rating Schedules<sup>23</sup>

					Score
25. Is he even-tempered or moody?					
Stolid, Rare changes of mood (3)	Generally very even- tempered (1)	Is happy or depressed as conditions warrant (2)	Strong and frequent changes of mood (4)	Has periods of extreme elations or depressions (5)	_____
26. Is he easily discouraged or is he persistent?					
Melts before slight obstacles or objections (5)	Gives up before adequate trial (3)	Gives everything a fair trial (1)	Persists until convinced of mistake (2)	Never gives in, Obstinate (4)	_____
27. Is he generally depressed or cheerful?					
Dejected, Melancholic, In the dumps (3)	Generally dispirited (4)	Usually in good humor (1)	Cheerful, Animated, Chirping (2)	Hilarious (5)	_____

for different items, indicate in descending order the relative desirability of the stated condition.

Another type of instrument that is in effect a rating scale is the Baker "Telling What I Do" tests. The accompanying illustration from the advanced level test for pupils in Grades 7 to 9 illustrates

Excerpt from Baker "Telling What I Do" Test<sup>24</sup>

On this sheet you will find many things about yourself. Some of these things are known about you already, but we want you to tell us yourself.

Each exercise has three answers. You are to draw a line under the one answer to each exercise that most nearly tells what you do. Put the letter of the answer in the parenthesis at the end of the line.

Underline only one answer to each exercise. Take the one that most nearly fits you. Be honest with yourself. Underline what you really do, even if it is not what you know you should do.

There are eighty exercises. Answer all of them. Take your time, and think over each exercise carefully. It should take you at least half an hour, or longer, to do all the exercises as you really should.

- |                          |                          |                              |
|--------------------------|--------------------------|------------------------------|
| 1. Tardy for school      |                          |                              |
| a. Never tardy           | b. Often tardy           | c. Tardy once in a while ( ) |
| 2. When I lose a game    |                          |                              |
| a. I just quit           | b. Don't care if I lose  | c. Try harder next time ( )  |
| 3. Eating                |                          |                              |
| a. Usually hurry         | b. Eat very fast         | c. Eat slowly ( )            |
| 4. When I meet strangers |                          |                              |
| a. Like to meet them     | b. Don't care about them | c. They bore me ( )          |
| 5. If I borrow           |                          |                              |
| a. I never pay back      | b. Pay back right away   | c. Pay when asked ( )        |

<sup>23</sup> M. E. Haggerty, W. C. Olson, and E. K. Wickman, *Haggerty-Olson-Wickman Behavior Rating Schedules*. Published by World Book Co., 1930.

<sup>24</sup> Harry J. Baker, "Telling What I Do," Advanced Form. Published by Public School Publishing Co., 1930.



the method of measuring pupil behavior. Scores can be obtained for the following areas of behavior: (1) school, (2) home, (3) play, (4) social, and (5) ethical-moral.

The *Hayes Scale for Evaluating the School Behavior of Pupils Ten to Fifteen* is another rating scale used either by the teacher in rating his pupils or by the pupils in obtaining self-ratings. Directions for using the scale and a few sample items are given in the accompanying illustration. The results of a simple scoring procedure can be used in preparing a school behavior profile, in locating maladjusted pupils, and in spotting the behavior areas in which maladjustment seems to exist.

### Excerpt from Hayes Scale for Evaluating School Behavior <sup>25</sup>

#### Directions for Using this Scale

Following is a list of habits which children 10 to 15 years old have been found to show. No one child could have all the habits listed, but is certain to have a considerable number of them.

Draw a circle around the T, F or U before each item to indicate: (T) you believe the statement is true of the child being rated; (F) you believe the statement is not true of the child being rated; (U) you are uncertain whether the statement is true or not true of the child being rated. Be sure to draw a circle about *one* letter and *one* only for every item in the list. Two samples are given below:

- (T) F U usually accepts responsibility when the occasion arises  
T (F) U often wastes time

Circle the following items in a similar manner

#### I

- |   |   |   |  |
|---|---|---|--|
| T | F | U | 1. often does little things to make others happy               |
| T | F | U | 2. usually thinks of consequences both to self and others      |
| T | F | U | 3. usually accepts responsibility when the occasion arises     |
| T | F | U | 4. often shares with others                                    |
| T | F | U | 5. usually does his share in any group activity                |
| T | F | U | 6. often "plays hookey" from school                            |
| T | F | U | 7. usually does the work expected of him                       |
| T | F | U | 8. usually defends his friends only when they are in the right |
| T | F | U | 9. usually makes friends easily                                |
| T | F | U | 10. often starts fights  |
| T | F | U | 11. usually quickly forgives wrongs done to him                |
| T | F | U | 12. often uses vulgar or profane words                         |
| T | F | U | 13. usually eats lunch with a group                            |

<sup>25</sup> Margaret Hayes, *A Scale for Evaluating the School Behavior of Children Ten to Fifteen*. Published by Psychological Corporation, 1933.

## 6 EVALUATIVE TECHNIQUES

The major instruments discussed above for use in the measurement of personality are paper-and-pencil instruments. They have come to be known as structured inventories because persons filling them out respond to the content of the instruments. In contrast are the unstructured techniques for evaluating personality. Although the various projective methods are most commonly referred to as unstructured, certain other evaluation techniques also permit free responses by the pupil. The unstructured techniques differ from the structured inventories in their direct concern with overt behavior of the whole child rather than with verbalized responses to specific situations. The two types of unstructured techniques dealt with below are used in the evaluation of individual pupils and the dynamics of group behavior.

### Evaluation of individual behavior

Three evaluative techniques useful in the study of individual pupils are considered briefly below. The anecdotal record and the case study are appropriately used by classroom teachers, but projective techniques should be employed only by psychological examiners, school psychologists, clinical psychologists, or other persons with technical training in their use.

*Anecdotal records.* Teachers have doubtless for generations used the anecdotal method in their spare-time discussions about pupils. However, its first use as an evaluative instrument was probably as recent as 1928.<sup>26</sup> The anecdotal record is an objective description by the teacher of a significant occurrence or episode in the life of the pupil. Unless a situation has sufficient meaning to a teacher who is alert to the underlying motives governing human behavior to bring it definitely to his attention, it probably is not of sufficient significance for inclusion in the anecdotal record.

An anecdotal record must be carefully, although not laboriously, prepared if it is to be of value. The anecdote is a highly objective brief of what occurred in a situation in which a pupil behaved in a sufficiently unusual manner to make the incident meaningful. It

<sup>26</sup> D. A. Robertson, chairman, "Report of Subcommittee on Personality Measurement." *Educational Record*, 9:53-68, Supplement No. 8; July 1928.



may consist of an objective narrative of the incident only or it may consist of the narrative, an impartial interpretation of the occurrence, and, as a possible third stage, even a recommendation for guidance of the pupil concerned. If interpretations and recommendations are given, however, they should be distinguished from the original description so that their nature is clearly apparent to a person reading the anecdotal record. The anecdotal record has great value only when it is made cumulative by the addition of new anecdotes as meaningful situations arise and are observed and recorded by the teacher or some other school officer.

*Case study.* The case study is a broad and comprehensive approach to the problems of pupil behavior. It should include extensive information about the present status of the pupil as well as about his past experiences and his family background. In fact, the case study may well draw upon many or even all of the types of information contained in adequate cumulative pupil records.

Usually there is a specific reason for making a case study. Such an approach may be used to gain a better understanding of a failing pupil, or a pupil who is poorly adjusted in one or another of many possible ways.

*Projective techniques.* A simple characterization of projective techniques is that they attempt to induce the child to reveal his personality through his free responses to situations that can be observed by the psychologist. Bell stated that the "*purpose* of projective techniques is to gain insight into the individual personality," and that their method is "to reveal the total personality, or aspects of the personality in their framework of the whole."<sup>27</sup> Techniques classified as projective differ widely in the materials used, the methods of presentation to the pupil, and the methods of interpreting the pupil's behavior, but all are intended to bring forth behavior representative of the inner personality and to permit the psychologist to draw inferences concerning intrinsic motives.

Probably most widely used in this area are the *Rorschach* test and the *Thematic Apperception Test*. The *Rorschach* makes use of pupil interpretations of inkblots and the *TAT* employs a wide variety of pictures as the basis for pupil responses. The *Szondi Test* employs photographs of persons and uses pupil indications of likes and dislikes as the basis for analysis. Some of the other projective techniques

<sup>27</sup> John E. Bell, *Projective Techniques: A Dynamic Approach to the Study of the Personality*. Longmans, Green and Co., New York, 1948. p. 4.

involve drawing or painting, play, handwriting, the completion of pictures, and dramatic productions. Sims stated that the essay examination is a projective technique under some conditions.<sup>28</sup>

As only a trained psychologist should attempt to employ these projective techniques, this brief discussion is intended to familiarize the student with the general nature of a few of the most widely used projective methods. Teachers occasionally may encounter situations in which maladjusted pupils are studied by the use of these techniques, so they should be sufficiently familiar with the general procedures involved to be intelligent users in subsequent pupil guidance of the interpretations made by the examining psychologists.

## Evaluation of group dynamics

Two methods are now used quite widely in studying the behavior of the whole child in settings involving interactions among members of social groups. Information is thereby obtained concerning the place of the individual within the group and concerning group behavior as influenced by the contributions of the individual members. The methods briefly discussed below involve the use of the sociogram and of analyses of group interactions. Both have been influenced considerably by developments in the field of sociology.

*The sociogram.* When groups of individuals are thrown together, as in a grade group of pupils in the elementary school or a homeroom group or class in the high school, some type of social relationship inevitably exists between each pupil and every other pupil individually. The possible range of relationships is from that involved in very close friendships to that of rejection. However, the variety of social situations is so wide that a pupil who is rejected by another in a particular social situation may be sought out in another, and quite different, social framework. For example, a boy preferred by a certain teammate as captain of the football team might be rejected by the same teammate as a member of a debating team.

The sociometric method which leads to the production of a sociogram as the end product is quite simple to apply. Most typically each pupil in the group is asked to name his first, second, and perhaps third choices among other members of the group in several significant and pertinent types of social settings. Questions asking for the

<sup>28</sup> Verner M. Sims, "The Essay Examination Is a Projective Technique." *Educational and Psychological Measurement*, 8:15-31; Spring 1948.



expression of individual preferences for class president, the occupant of an adjacent seat in the homeroom, or a member of a committee merely illustrate the wide range of possibilities.

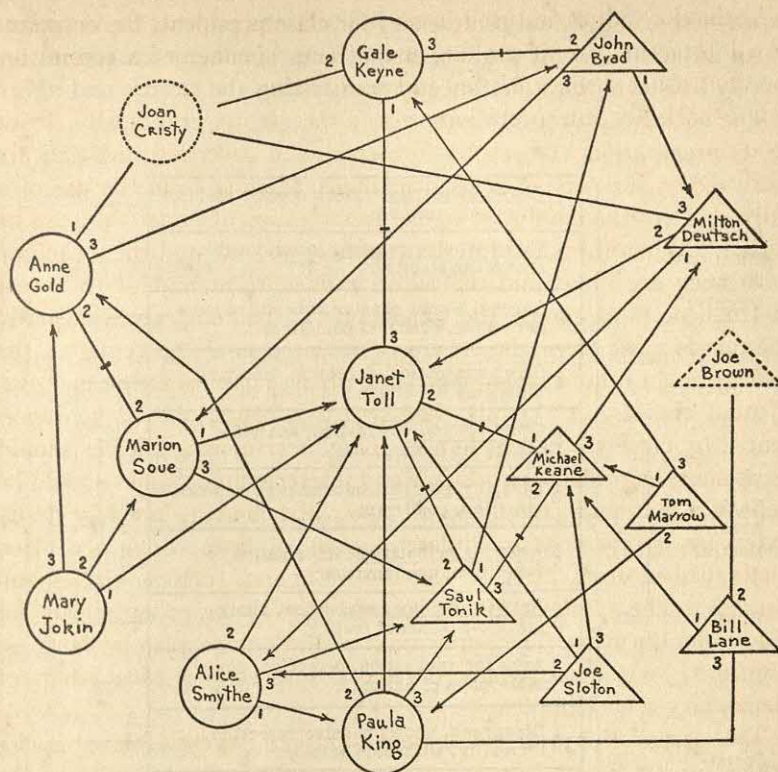
The sociogram is used to represent the results graphically. Prior to its preparation, the results in response to a certain question are analyzed by any one of several methods, ranging from the use of a tally sheet to the employment of cards or slips of paper that can be sorted.<sup>29</sup> When these results, showing first, second, and third choices, have been organized and the pupils ranked from highest to lowest in frequencies of choice, the sociogram can be constructed. Usually the pupils most often chosen are represented near the center of the sociogram and those rejected or least often chosen near the margins. Mutual choices, i.e., pupils choosing each other, should be represented by closely adjacent figures. The lines showing choices should be as short as reasonably possible and intercrossing of lines should be kept at a minimum. First, second, and third choices should be designated by numbers or by different types of lines. Boys are often distinguished from girls by the use of figures such as circles and triangles. The symbols should contain pupils' names or initials for ready identification. The application of these principles is shown in Figure 21, which represents the social interactions of a group of elementary-school pupils.

The preparation and subsequent study of several related sociograms for a class or homeroom group should add materially to the teacher's understanding of social relationships among his pupils and enable him to use the results in significant ways to take into account the relationships found and to take remedial action where conditions warrant.

*Direct observation.* The interactions of the individual members of a small group working on a common problem can be evaluated by the use of observational techniques. Groups should be small, say not larger than twenty, and should be working on cooperative projects in which interaction rather than individual work is entailed.<sup>30</sup> Although interaction process analysis is not a distinctly new technique, its methodology has been improved materially during the last ten years.

<sup>29</sup> Helen H. Jennings, *Sociometry in Group Relations: A Work Guide for Teachers*. American Council on Education, Washington, D. C., 1948. p. 17-21.

<sup>30</sup> Robert F. Bales, *Interaction Process Analysis: A Method for the Study of Small Groups*. Addison-Wesley Press, Inc., Cambridge, Mass., 1950. p. i.

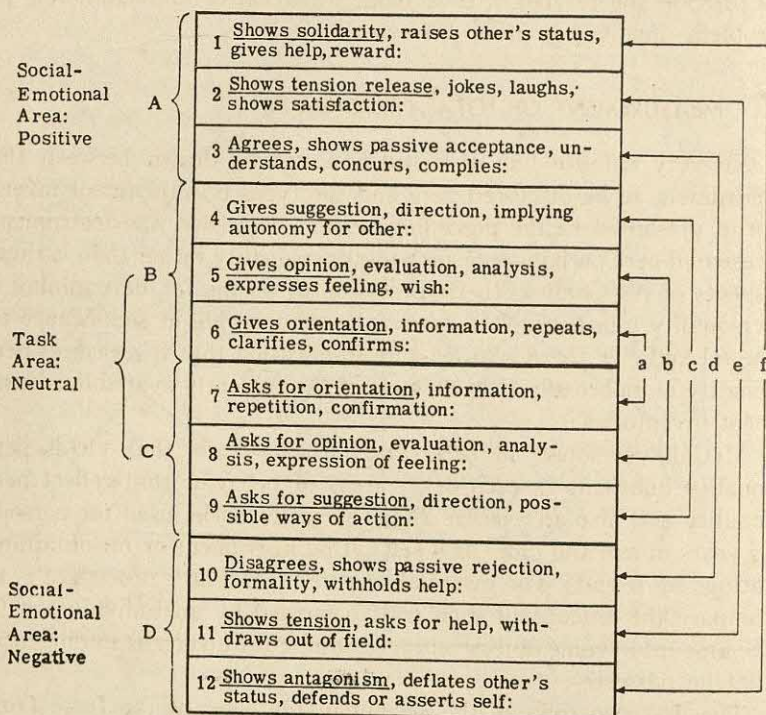
Fig. 21. Sample sociogram <sup>31</sup>

The observer, in studying group interactions, records in sequence the actions of the group members and classifies each item of observed behavior by interpreting it as belonging to one or another of the behavior types decided upon in advance as significant. An illustration

<sup>31</sup> Jennings, *op. cit.* p. 22.



of behavior categories used in such situations appears in Figure 22. An interaction profile making use of these behavior categories and various other devices is used in summarizing the results and affording a basis for interpretation.



## KEY:

- a Problems of Communication
- b Problems of Evaluation
- c Problems of Control
- d Problems of Decision
- e Problems of Tension Reduction
- f Problems of Reintegration
- A Positive Reactions
- B Attempted Answers
- C Questions
- D Negative Reactions

Fig. 22. Behavioral categories and their major relations <sup>32</sup>

<sup>32</sup> Bales, *op. cit.* p. 9.

As is true for projective methods, the teacher should not attempt to make direct use of these techniques for studying group interactions. However, sufficient insight into the purposes and general methods involved is probably given by the list of behavior categories to provide the teacher with a basis for a better understanding of problems involved in group behavior.

## 7 MEASUREMENT OF TOTAL PERSONALITY

No very definite line of distinction can be drawn between the instrument to be discussed here and the types of adjustment inventories presented in the preceding section. However, the instrument presented here perhaps measures total personality rather than various aspects of personality. In its provision of means for determining a personality quotient ( $PQ$ ), essentially comparable in significance to the  $IQ$  and  $EQ$ , there is at least an implication that it measures personality more broadly than do most of the currently available adjustment inventories.

McCall developed an *Inter-Trait Rating Scale* that yields personality quotients in each of 43 areas of behavior that reflect personality and also an average  $PQ$ . The scale can be used for persons 12 years of age and older as a self-rating instrument or for obtaining ratings by friends. The essential feature of McCall's procedure is to compare the amount of each trait possessed by an individual with the amount of some objectively measurable trait, such as intelligence, that he possesses.

The first two rows of the accompanying copy of the *Inter-Trait Rating Scale* are filled in with ratings for an hypothetical individual who has an  $IQ$  of 115, as a basis for showing how the scale is used. Because the rater feels that the individual is lower in accuracy than in intelligence, he places a "—" in the second column. He estimates that his judgment on the individual's accuracy is about 40 per cent of certainty. Consequently, he writes "40" in the third column. The  $PQ$  column is then filled out by taking half of the percentage of certainty and, because the sign in the second column is negative, subtracting it from the  $IQ$  of 115 to obtain a  $PQ$  of 95. On adaptability, the plus rating with a 30 per cent degree of certainty indicates that 15 should be added to the  $IQ$  of 115, to net a  $PQ$  of 130. The average of the 43 separate  $PQ$ s then becomes the general  $PQ$  for the individual.



McCall Inter-Trait Rating Scale <sup>33</sup>

Traits	Above or Below Intelligence	Percent of Certainty	Personality Quotients ( $\frac{1}{2}$ the % plus IQ)
Accuracy .....	—	40	95
Adaptability .....	+	30	130
Appearance .....			
Cheerfulness .....			
Conscientiousness .....			
Cooperativeness .....			
Courage .....			
Courtesy .....			
Decisiveness .....			
Democracy .....			
Effectiveness .....			
Enthusiasm .....			
Foresight .....			
Generosity .....			
Happiness .....			
Healthiness .....			
Independence .....			
Industriousness .....			
Initiative .....			
Leadership .....			
Likeableness .....			
Loyalty .....			
Open-Mindedness .....			
Orderliness .....			
Originality .....			
Persistence .....			
Pleasing Voice .....			
Poise .....			
Progressiveness .....			
Punctuality .....			
Refinement .....			
Reliability .....			
Self-Confidence .....			
Self-Control .....			
Sense of Humor .....			
Sincerity .....			
Sociability .....			
Sympathy .....			
Tact .....			
Thoroughness .....			
Tolerance .....			
Truthfulness .....			
Vivacity .....			

<sup>33</sup> William A. McCall, *Measurement*. Macmillan Co., New York, 1939. p. 315.

McCall stated that the embarrassment that sometimes arises when one person is asked to rate a friend in the friend's presence will not arise with this rating scale. He said, in commenting upon the manner in which some friends rated him: <sup>34</sup>

Since they could not rate him down in accuracy without rating him up in intelligence or up in adaptability without rating him down in intelligence there was no particular embarrassment to them or him in these ratings, although the author does not see himself as others see him at certain points. They were not asked to state whether the author was very dull or very intelligent or very accurate or inaccurate, nor even to state how much difference there is between his intelligence and his accuracy.

### Topics for Discussion

1. In what way is a knowledge of personality measurement procedures valuable to the teacher?
2. What is meant by personality? How do psychologists and laymen differ in their conceptions of personality?
3. Briefly characterize two association methods of evaluating behavior.
4. Indicate the nature of observation procedures for the evaluation of personality.
5. What is the nature of graphic rating scales?
6. How are personal reports used in personality measurement?
7. Briefly indicate the nature of attitudes. Of what concern are they to the teacher?
8. Indicate the nature of one or two attitudes scales for use in the elementary or secondary school.
9. What is the nature of interests? How are pupil interests of significance to the teacher?
10. Discuss the two major procedures used in the measurement of interests.
11. What are the causes and symptoms of emotional maladjustment? Which are easier to recognize? Why?
12. Indicate some of the methods by which pupil adjustment is measured.
13. What are three major methods of evaluating individual behavior?
14. How should the teacher expect to be involved in the administration and use of results from projective techniques?
15. What are some of the modern methods for evaluating group dynamics?
16. In what ways can the classroom teacher appropriately use sociograms?

<sup>34</sup> *Ibid.* p. 314.



17. Discuss how total personality is measured by one technique. What is the PQ?

### Selected References

- ANDERSON, HAROLD H., AND ANDERSON, GLADYS L., editors. *An Introduction to Projective Techniques*. New York: Prentice-Hall, Inc., 1951.
- BELL, JOHN E. *Projective Techniques: A Dynamic Approach to the Study of the Personality*. New York: Longmans, Green and Co., 1948. Chapters 2-24.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 67-293, 726-51.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p. 49-100.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938. p. 53-62.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. p. 51-218.
- CATTELL, RAYMOND B. *An Introduction to Personality Study*. New York: Longmans, Green and Co., 1950.
- CATTELL, RAYMOND B. *Description and Measurement of Personality*. Yonkers, N. Y.: World Book Co., 1946.
- DUNKEL, HAROLD B. *General Education in the Humanities*. Washington, D. C.: American Council on Education, 1947. Chapters 2-3; p. 267-96.
- FERGUSON, LEONARD W. *Personality Measurement*. New York: McGraw-Hill Book Co., Inc., 1952.
- FRANK, LAWRENCE K. *Projective Methods*. Springfield, Ill.: Charles C. Thomas, 1948.
- FREEMAN, FRANK N. *Mental Tests: Their History, Principles and Applications*. Revised edition. Boston: Houghton Mifflin Co., 1939. Chapter 8.
- FREEMAN, FRANK S. *Theory and Practice of Psychological Testing*. New York: Henry Holt and Co., 1950. Chapters 13-14.
- FROELICH, CLIFFORD P., AND DARLEY, JOHN G. *Studying Students: Guidance Methods of Individual Analysis*. Chicago: Science Research Associates, 1952. Chapters 13-14.
- FRYER, DOUGLAS. *The Measurement of Interests*. New York: Henry Holt and Co., 1931.
- GOODENOUGH, FLORENCE L. *Mental Testing: Its History, Principles, and Applications*. New York: Rinehart and Co., Inc., 1949. Chapter 27.

- GREENE, EDWARD B. *Measurements of Human Behavior*. Revised edition. New York: Odyssey Press, 1952. Chapters 20-24.
- HARSH, C. M., AND SCHRICKEL, H. G. *Personality Development and Assessment*. New York: Ronald Press Co., 1950.
- HARTSHORNE, HUGH, AND MAY, MARK A. *Studies in Deceit*. New York: Macmillan Co., 1928.
- HAVIGHURST, ROBERT J., AND TABA, HILDA. *Adolescent Character and Personality*. New York: John Wiley and Sons, Inc., 1949. Part 5.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. Chapters 16-18.
- MALLER, JULIUS B. "Personality Tests." *Personality and the Behavior Disorders*. New York: Ronald Press Co., 1944. Chapter 5.
- MURSELL, JAMES L. *Psychological Testing*. Second edition. New York: Longmans, Green and Co., 1949. Chapter 8.
- OLSON, WILLARD C. "Personality." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 806-17.
- SARGENT, HELEN. "Projective Methods: Their Origin, Theory, and Application in Personality Research." *Psychological Bulletin*, 42:257-93; May 1945.
- SHEVIAKOV, GEORGE V., AND BLOCK, JEAN F. "Evaluation of Personal and Social Adjustment." *Appraising and Recording Student Progress*. New York: Harper and Brothers, 1942. Chapter 6.
- SHEVIAKOV, GEORGE V., AND FRIEDBERG, JEAN. "The Use of Interest Inventories for Personality Study." *Journal of Educational Research*, 33:692-97; May 1940.
- STAGNER, ROSS. "Attitudes." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 77-84.
- SUPER, DONALD E. *Appraising Vocational Fitness by Means of Psychological Tests*. New York: Harper and Brothers, 1949. Chapters 16-19.
- SYMONDS, PERCIVAL M. *Diagnosing Personality and Conduct*. New York: D. Appleton-Century Co., Inc., 1931.
- SYMONDS, PERCIVAL M. "Development and Educational Significance of Projective Technics in Personality Measurement." *Review of Educational Research*, 20:51-62; February 1950.
- TRAXLER, ARTHUR E., AND JACOBS, ROBERT. "Construction and Educational Significance of Structured Inventories in Personality Measurement." *Review of Educational Research*, 20:38-50; February 1950.
- TRAXLER, ARTHUR E. *The Use of Tests and Rating Devices in the Appraisal of Personality*. Educational Records Bulletin No. 23. New York: Educational Records Bureau, March 1938.



## ***Summarizing the Results of Measurement***

THE FOLLOWING points in the summarization of test results are considered in this chapter :

- A. Statistical procedures in summarizing test results.
- B. Tabulation of test scores.
- C. Common measures that express typical performance.
- D. Common measures of spread or variability.

It is common knowledge that a wide range of ability may be expected from the different individuals in a given class and that scores representing objective measures of achievement or other traits will vary widely. Since the human mind is not able to grasp and hold numerous unlike facts in isolation, accurate description of test results depends upon their statistical summarization. Summaries and descriptions of this type need not disturb the student, for after all most of these elementary statistical procedures are simple. Actually, the main requirements are the learning of a new and different type of vocabulary and the use of a few relatively simple arithmetic skills.

The use of statistical methods in the analysis of test results is directly in line with good scientific technique. Scientific method in handling test results involves :

- (1) *The collection of facts.* Within the limits of accuracy of the tests used, the test scores may be said to represent facts.

- (2) *The classification and organization of the facts.* Simple statistical practices of grouping and tabulating data are utilized for this purpose.
- (3) *The further reduction and analysis of the data.* Such common statistical procedures as determining measures of central tendency, variability, and relationship are required at this point.
- (4) *The interpretation of test data.* Graphical representations and various types of derived scores are involved here.
- (5) *The validation of tests.* Procedures for determining the validity, reliability, and objectivity of tests illustrate this need.

The most important statistical techniques from the standpoint of the frequency of their use in education involve abilities to: (1) classify and tabulate data, (2) determine and apply the common measures of central tendency, (3) determine and apply the common measures of spread or variability, (4) secure derived scores and use them in the interpretation of test results, (5) utilize graphical methods in the presentation and interpretation of test results, and (6) apply correlational procedures in determining the relationship between two sets of data. The discussion and explanation of these techniques constitute the major portions of this and the two following chapters.

This relatively large amount of emphasis is given to these points for two reasons: (1) Successful and satisfactory work with test results can be expected only when the person using them is adequately equipped to understand and interpret them. Such abilities are dependent on a reasonable mastery of these elementary statistical techniques. (2) Current educational literature in practically all fields is literally filled with the terms and the techniques discussed here. Reports of progress in education are dependent on statistical methods. If the teacher and the student are to keep up to date educationally, they must develop the ability to read with understanding the practical aspects of statistical discussions in current educational literature.

## 1 CLASSIFICATION AND TABULATION OF TEST SCORES

### Need for a method of grouping data

The very fact that people are unlike physically and mentally gives rise to the need for statistical methods in psychology and education. For example, it may be observed readily from Table 9 that there



are great differences in the scores made by the thirty-seven pupils who took a certain reading test. However, it requires rather careful scrutiny to determine that the highest and lowest scores are respectively 72 and 24, while very little further information can be obtained from these scores without rearranging them.

TABLE 9. Reading test scores of 37 ninth-grade pupils in alphabetical order of last names

72	45	57	70	47	51	34	32
46	58	24	65	42	36	55	46
63	45	46	52	40	48	48	
48	68	50	49	43	60	55	
49	43	40	60	50	54	30	

The relatively simple practice of arranging test scores in order of size from highest to lowest is helpful, however. Table 10 reproduces the reading test scores of the thirty-seven pupils in descending order. It now is more easily apparent than from Table 9 that the highest and lowest scores are respectively 72 and 24, while it can also rather easily be determined that the middle score, or *midscore*, is 48.

TABLE 10. Reading test scores of 37 ninth-grade pupils in descending order

72	60	55	50	48	45	40	30
70	60	54	49	47	45	40	24
68	58	52	49	46	43	36	
65	57	51	48	46	43	34	
63	55	50	48	46	42	32	

Four consistent ways in which these same thirty-seven scores can be classified into a *frequency distribution* are shown in Table 11. The first illustration, in which the scores retain their individual identities, may be called a *simple or ungrouped frequency distribution*. The other three illustrations, in which the grouping of scores destroys the individual identities of most of them, are called *grouped frequency distributions*. The first distribution furnishes the basis for obtaining detailed information concerning these scores, but such information would be rather costly in the time required to derive it. The fourth illustration furnishes the basis for obtaining quick but quite unsatisfactory information, for the very rough grouping almost

entirely sacrifices even the approximate identity of the individual scores. The second and third illustrations, neither of which demands an undue time expenditure in order to obtain accuracy nor sacrifices accuracy for a saving in time and labor, represent satisfactory practices midway between the two extreme methods of handling the scores. The second is somewhat preferable to the third for these data.

TABLE II. Reading test scores of 37 ninth-grade pupils in frequency distributions

Intervals of 1 Unit						Intervals of 3 Units		Intervals of 5 Units	
Scores	f	Scores	f	Scores	f	Scores	f	Scores	f
72	1	55	2	39		71-73	1	68-72	3
71		54	1	38		68-70	2	63-67	2
70	1	53		37		65-67	1	58-62	3
69		52	1	36	1	62-64	1	53-57	4
68	1	51	1	35		59-61	2	48-52	9
67		50	2	34	1	56-58	2	43-47	8
66	1	49	2	33		53-55	3	38-42	3
65		48	3	32	1	50-52	4	33-37	2
64		47	1	31		47-49	6	28-32	1
63	1	46	3	30	1	44-46	5	23-27	1
62		45	2	29		41-43	3	Intervals of 15 Units	
61		44		28		38-40	2		
60	2	43	2	27		35-37	1	Scores	f
59		42	1	26		32-34	2	68-82	3
58	1	41		25		29-31	1	53-67	9
57	1	40	2	24	1	26-28		38-52	20
56						23-25	1	23-37	5

In the preparation of a frequency distribution, the method of grouping test scores is not dissimilar to that followed by postal clerks in the distribution of outgoing letters in a large railway postal terminal. Mail designated for certain sections of the country or for certain large centers from which it is redistributed is thrown into the proper mail pouches. Some pouches, however, contain mail addressed to a number of post offices in the same section of the country. For example, in the Chicago postal terminal, mail addressed to post offices in the Pacific northwest may be consigned to a certain group



of pouches, while mail going to the southwest section of the United States will be consigned to other pouches representing that section of the country. The number of pouches required depends on the number of pieces of mail to be distributed and also on the population of the section of the country that can be most efficiently served by a given pouch. Increasing the number of pouches naturally increases the labor involved in sorting the pieces of mail, but at the same time it increases the accuracy of the distribution. Mail in Chicago might be sorted into two classes—eastbound and westbound. This would introduce a large error, since not all sections of the country would be effectively served by this rough classification. The other extreme, of using at this point a separate pouch for the mail addressed to each post office, would be entirely impracticable.

### Classifying and tabulating scores

The foregoing illustration will give the reader a clear conception of the purposes and the problems involved in grouping test scores into a frequency table. The steps of procedure presented in the following paragraphs are for the student's guidance in the preparation of grouped frequency distributions of test scores.

(1) *Determine the range.* Find the highest score and the lowest score in the series. Find the difference between these scores. The difference is called the *range* (*R*) of the distribution. In the case of the scores given in Table 9, the range is  $72 - 24$ , or 48. The range is useful in determining the number of class intervals to use in the frequency table.

(2) *Determine the size of the class intervals.* Use the range to determine whether the scores should be classified by units of 1, 3, 5, 7, or 15, i.e., to determine the size of the class intervals. For the range of 48 found for the scores of Table 9, a practical grouping is by intervals of 3 units each.

No special rule relative to the number of class intervals to be used in a frequency table can be stated. However, it is usually unwise to group data into fewer than ten to twelve class intervals because of the greater *error of grouping* as the number of intervals is decreased. Likewise, it is usually undesirable to use more than twenty to twenty-five class intervals because of the increased labor involved. The main idea of grouping the scores into approximately twelve to twenty or

so intervals is to classify the scores into a sufficiently small number of groups that they may be thought about effectively and yet not into so few groups that important differences are covered up or significant errors are introduced.

TABLE 12. Relation between range of scores and size of class intervals

For a Range of	Use a Class Interval of
25 or less	1
26 to 69	3
70 to 125	5
126 to 175	7
176 or more	15

(3) *Set up the frequency table.* Construct the table into which the scores are to be tabulated by use of the following steps of procedure:

(a) *Label three column headings c.i., Tabulation, and f.* The first and third headings are commonly-used abbreviations for "class interval" and "frequency" respectively.

(b) *Determine the limits of the highest class interval.* First find the multiple of the class interval that is closest to or equal to the highest score in the series. This number is the midpoint of the highest class interval. Then establish the integral, i.e., whole-number, limits of the interval equal distances above and below this midpoint, so that the distance between the *integral limits* is one scale unit less than the size of the class interval. The *real limits* of the intervals are then found by continuing upward .5 of a score unit above the higher integral limit and .5 of a score unit below the lower integral limit. Figure 23 shows how the midpoint, integral limits, and real limits appear for this highest class interval.

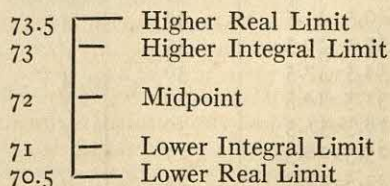


Fig. 23. Midpoints and limits of a class interval



(c) *Write the integral limits of the class intervals in the c.i. column.* Start at the top with the interval that will include the highest score and continue downward consistently to include at the bottom the interval that will include the lowest score.

At this point it may help to refer again to the scores given in Table 9. For these data it was determined above that the grouping should be by class intervals of 3 units each. Since the highest score, 72, is exactly divisible by 3, 72 is the midpoint of the highest class interval. Then 71 and 73 will become the lower and higher integral limits respectively, and 70.5 and 73.5 will become the lower and higher real limits respectively, of the class interval. The next lower interval will have a midpoint of 69, integral limits of 68 and 70, and real limits of 67.5 and 70.5. The first three columns of Table 13 show these various points for each interval for the entire distribution based on the scores of Table 9.

TABLE 13. Reading test scores of 37 ninth-grade pupils in a grouped frequency distribution

Class Interval (c.i.)			Tabulation	Frequency (f)
Integral Limits	Real Limits	Mid- point		
71-73	70.5-73.5	72	/	1
68-70	67.5-70.5	69	//	2
65-67	64.5-67.5	66	/	1
62-64	61.5-64.5	63	/	1
59-61	58.5-61.5	60	//	2
56-58	55.5-58.5	57	//	2
53-55	52.5-55.5	54	///	3
50-52	49.5-52.5	51	////	4
47-49	46.5-49.5	48	//// /	6
44-46	43.5-46.5	45	////	5
41-43	40.5-43.5	42	////	3
38-40	37.5-40.5	39	//	2
35-37	34.5-37.5	36	/	1
32-34	31.5-34.5	33	//	2
29-31	28.5-31.5	30	/	1
26-28	25.5-28.5	27		0
23-25	22.5-25.5	24	/	1
				$N = 37$

The above directions and illustration will be clarified if the terms are reviewed and defined. The *class interval* is the group, or compartment, within the limits of which given scores are assigned. The *midpoint* is a point midway between the upper and the lower limits of the interval. The *integral limits* are the limits or boundaries of the interval in terms of whole numbers. The *real limits* are the actual boundaries of the interval. For convenience in tabulation it is found desirable to choose the limits of the interval in such a way that the midpoint is a whole number. This, of course, makes it necessary that the upper and lower real limits be fractional values whenever odd-sized intervals are used. Many statisticians prefer this method because they recognize that, although test scores are usually not given in fractional values, a score of a certain value, say 72, might, if the measurement were more accurate, equally well represent a score a fraction above 72 or a fraction below 72. A score of 72, then, represents any score between 71.5 and 72.5. This method has the merit of furnishing a natural location on the scale for all scores expressed in whole numbers.<sup>1</sup>

(4) *Tabulate the scores.* Begin with the first score in the original list of scores. Determine in which class interval this score will be included. Place a tally mark in the *Tabulation* column of the appropriate class interval. Make another tally mark for the second score in the interval in which it is included. Continue thus until a tally mark has been made for each score in the series. Make each fifth mark in any interval a slanting mark across the preceding four tally marks. Complete the frequency distribution by totaling the tally marks in each row and writing the proper number for that row

<sup>1</sup> This represents one of the two most common assumptions made in the statistical work concerning the meaning of a test score. The other widely used statistical method assumes that the true score, of which the test score actually obtained is only an estimate, is not likely to be less than the obtained score but may lie anywhere between the obtained score and a score one unit greater. For example, a score of 72 represents a true score somewhere from 72 to 72.9999....

The authors believe that the method used in this volume represents a more defensible assumption concerning the meaning of a test score. However, instructors preferring to use the other method can do so easily by shifting each midpoint and each real limit of a class interval .5 of a score point upward from the values given in this and the two following chapters. For example, the real limits of 70.5 and 73.5 for the interval 71-73 would become 71.0 and 73.9999... in the other method, and the midpoint of 72 for the same interval would become 72.5 in the other method. It should be noted, also, that this procedure results in an arithmetic mean that is exactly .5 larger than for the method used in this volume.



in the  $f$  column, and then by obtaining the sum of these frequencies. This sum,  $N$ , should equal the total number of original scores.

### Summary of steps in classifying and tabulating scores

The classroom teacher will sometimes find the construction of a frequency table unnecessary in his experience with tests, since he often works with small numbers of cases and can check the scores from the papers themselves. However, there are many occasions when the frequency distribution is necessary. It is an effective way of recording and preserving the results of using tests in the classroom. It makes possible a number of short cuts in the calculation of certain statistical measures useful in interpreting test results. The methods by which these measures are computed are given in succeeding sections of this chapter. This section summarizes in concise form the steps of procedure in setting up a frequency table and in tabulating scores.

- (1) Determine the range. Find the highest score and the lowest score and obtain the difference between them. ( $R$ )
- (2) Determine the size of the class intervals. If the result of step (1) is: 25 or less, use a class interval of 1; 26 to 69, use a class interval of 3; 70 to 125, use a class interval of 5; 126 to 175, use a class interval of 7; and 176 or more, use a class interval of 15. ( $c.i.$ )
- (3) Set up the frequency table:
  - (a) Label three column headings  $c.i.$ , *Tabulation*, and  $f$ ,
  - (b) Determine the limits of the highest class interval so that its midpoint is divisible by, and the distance between its *real limits* is equal to, the size of the interval determined in step (2), and
  - (c) Write the integral limits of the class intervals in the  $c.i.$  column, starting at the top with the interval that contains the highest score and continuing downward consistently to include the interval that contains the lowest score.
- (4) Tabulate the scores. Place a tally mark in the *Tabulation* column opposite the  $c.i.$  that indicates the proper position for each score, carry across the total of the tally marks in each class interval to the  $f$  column, and add the frequencies in the  $f$  column to obtain the total number of cases. ( $N$ )

## Exercises in Tabulating Test Scores

1. Set up a frequency table and tabulate the arithmetic test scores listed below for 40 fifth-grade pupils.

59, 48, 38, 66, 57, 42, 51, 66, 75, 53, 55, 47, 41, 35, 49, 71, 63,  
55, 51, 44, 79, 66, 57, 58, 51, 45, 52, 50, 48, 72, 51, 54, 64, 59,  
53, 42, 53, 55, 58, 61.

2. Set up a frequency table and tabulate the language test scores listed below for 30 tenth-grade pupils.

56, 47, 60, 52, 39, 65, 41, 81, 69, 30, 21, 64, 44, 55, 28, 11, 6,  
24, 49, 45, 10, 49, 46, 39, 24, 64, 29, 34, 46, 34.

3. Set up a frequency table and tabulate the spelling test scores listed below for 30 ninth-grade pupils.

19, 14, 11, 9, 17, 15, 13, 13, 6, 16, 21, 10, 12, 18, 11, 13, 10, 5, 8,  
15, 14, 18, 4, 11, 12, 13, 22, 6, 21, 14.

## 2 MEASURES OF CENTRAL TENDENCY

### Need for measures of central tendency

The grouping of test scores into frequency tables is one step in the process of condensing them so that they can be analyzed and interpreted. However, a further step must be taken before it is possible to describe the data. Some single term or value that is representative of the entire table must be found. Since these values which may be taken to represent an entire distribution of scores are usually found near the center of the data when arranged in order of size, they are commonly called *measures of central tendency*. The three common measures of central tendency are: (1) the arithmetic mean, (2) the median, and (3) the mode. Of these three measures of central tendency, the median and the arithmetic mean are used almost exclusively in educational measurements and are, accordingly, the only ones emphasized in this discussion.

### Computing the arithmetic mean of ungrouped data

The arithmetic mean is the best known and the most widely used measure of central tendency. Indeed, the word "average" is thought by many persons to designate the arithmetic mean, although the arithmetic mean is only one of several "averages."



Practically everyone knows how to find and use the so-called average or arithmetic mean. It is commonly defined as the measure resulting from dividing the sum of the measures in the distribution by the number of measures. Thus the arithmetic mean of the scores 93, 90, 89, 88, and 86 is  $446 \div 5 = 89.2$ . The value of this measure lies in the fact that it lends itself to describing by means of a single term a group of widely varying scores or measures. It expresses in very compact form one specific fact about the scores in which each single score has a part. On this account it is one of the basic statistical measures of central tendency.

### Computing the arithmetic mean of grouped data

The arithmetic mean can also be readily found from a frequency distribution. In order to make the procedure somewhat more definite it is advisable to redefine the term arithmetic mean. When considered from this point of view *the arithmetic mean is defined as a point on the scale such that the sum of the deviations of the values larger exactly equals the sum of the deviations of the values smaller than it is*. Expressed in physical terms, it may be thought of as the point at which the fulcrum must be placed in order to balance the scale, when it is considered as a beam of varying thickness or density. This point may be determined experimentally or by mathematical calculation. Without regard to the method employed, the fulcrum must be so placed that the moments of the forces on one side are exactly equaled by the moments of the forces on the other side.

Figure 24 illustrates the principle of moments of force by a beam in balance when a weight of one pound is suspended three feet from the fulcrum and a weight of three pounds is suspended one foot from the fulcrum.

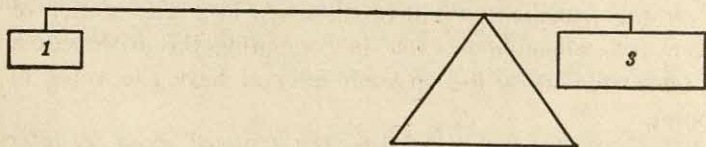


Fig. 24. Moments of force and the arithmetic mean

The parallel between the physical lever and the mathematical calculation of the arithmetic mean is quite close. The problem in each case is to balance the forces on either side of a point to be determined.

If the physical lever is out of balance, the correction is made by moving the fulcrum in the direction of the heavier end until equilibrium is established. In calculating the arithmetic mean a sort of trial balance is taken. If the moments of force are too great on one side, the point of rotation is similarly moved in the direction of the heavy end until the difference between these two forces becomes zero.

This may be aptly illustrated by a procedure which classroom teachers have undoubtedly frequently used. For example, it is desired to obtain the average of the scores of 93, 90, 89, 88, and 86. By inspection it may be seen that 89 is approximately the correct mean. The 90 is one point too large, the 93 is four points too large. In a corresponding way 88 is one point too small and 86 is three points too small. The total of the differences above the assumed mean is five and the total of the differences below the assumed mean is four; therefore the assumed mean of 89 is too small by the amount of this difference divided by the number of cases. Since this is equal to  $1 \div 5$ , or .2, the mean is 89.2. This checks exactly with the mean found by the method of totaling the measures and then dividing by the number of measures, given on page 318.

This method of computing the arithmetic mean will now be applied to the grouped frequency distribution given in Table 13 for thirty-seven reading test scores.

(1) *Assume a value for the mean.* The midpoint of a class interval near the middle of the frequency distribution should be taken as the assumed mean. This class interval is usually chosen so that it fairly closely approximates the arithmetic mean. As a matter of fact, however, the results will be the same regardless of the particular interval whose midpoint is chosen as the assumed mean. In the illustration of Table 14, it has been estimated that the arithmetic mean will fall in or near the interval 50-52, so the assumed mean is 51.00. For reasons that will be discussed in a later section of this chapter, it is common practice in computing the arithmetic mean to assume that all scores in each interval have the value of the midpoint.

(2) *Lay off the deviations from the assumed mean by intervals.* Fill in the *d* column by assigning a deviation of 0 to the class interval in which the assumed mean is located and then counting both upward and downward from that interval by units. Deviations above the assumed mean are positive and those below the assumed mean are negative. Since this is equivalent to showing the number of class



intervals by which each interval deviates from the one containing the assumed mean and also its direction from that interval, the deviations are said to be stated in terms of class intervals.

TABLE 14. Computation of the arithmetic mean for the grouped frequency distribution of 37 reading test scores

Class Interval ( <i>c.i.</i> )		Fre- quency ( <i>f</i> )	Devia- tion ( <i>d</i> )	<i>fd</i>	
Integral Limits	Mid- point				
71-73	72	1	+7	+ 7	1. Assume a value for the mean (51.00)
68-70	69	2	+6	+12	2. Lay off the deviations from the assumed mean by intervals
65-67	66	1	+5	+ 5	3. Add the <i>fd</i> column to the table
62-64	63	1	+4	+ 4	4a. Obtain the products of the frequencies and deviations by intervals
59-61	60	2	+3	+ 6	
56-58	57	2	+2	+ 4	4b. Determine the algebraic sum of the deviations of the scores
53-55	54	3	+1	+ 3	$\Sigma fd = +41 + (-66) = -25$
50-52	51	4	0	0	
47-49	48	6	-1	- 6	4c. Determine the mean of the deviations of the scores
44-46	45	5	-2	-10	$\Sigma fd / N = -25 / 37 = -.68$
41-43	42	3	-3	- 9	
38-40	39	2	-4	- 8	4d. Convert the mean of the deviations of the scores to the scale value <i>c.i.</i> $\times \Sigma fd / N = 3 \times$
35-37	36	1	-5	- 5	$-.68 = -2.04$
32-34	33	2	-6	-12	
29-31	30	1	-7	- 7	5. Obtain the arithmetic mean.
26-28	27	0	-8	0	Assumed mean + <i>c.i.</i> $\times$
23-25	24	1	-9	- 9	$\Sigma fd / N = 51.00 + (-2.04)$
		$N = 37$		$= 48.96$ (A.M.)	

(3) Add a column at the right of the table and label it *fd*. This column is used for recording the products of the frequencies and the corresponding deviations.

(4) Obtain the correction to the assumed mean. Obtain this correction by use of the following steps of procedure:

(a) Obtain the products of the frequencies and deviations by intervals. Multiply each frequency by its corresponding deviation and place the results in the *fd* column. Products below the assumed mean will have negative signs.

(b) *Determine the algebraic sum of the deviations of the scores.* Add the  $fd$  values algebraically to obtain  $\Sigma fd$  by determining the sum of the positive values, the sum of the negative values, and then assigning the sign of the larger value to their difference. This is equivalent to finding the magnitude of forces at one end and at the other end and then obtaining their difference, in the illustration of a beam resting on a fulcrum. Since for the distribution of Table 14, the positive  $fd$  values total 41 and the negative  $fd$  values total  $-66$ , their algebraic sum is  $-25$ . This is shown graphically in Figure 25, which illustrates the scores of the frequency distribution of Table 14 distributed along a beam resting on a fulcrum at the point of the assumed mean.

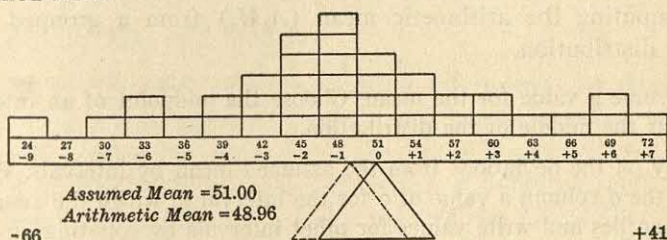


Fig. 25. Moments of force for the 37 reading test scores

(c) *Determine the mean of the deviations of the scores.* Divide  $\Sigma fd$  by  $N$  to obtain  $\Sigma fd/N$ , retaining the proper sign. In order to bring about an exact balance of these two forces, the fulcrum must be moved slightly in the direction of the heavier end of the scale, which is, in this case, in a minus direction. Since there are 37 measures in the distribution and each of them contributes equally to the resultant force of  $-25$  units, the average correction is the result of dividing  $-25$  by 37, or  $-.68$ .<sup>2</sup>

(d) *Convert the mean of the deviations of the scores to the scale value.* Since each interval in this table is three units in size, it is necessary to multiply  $-.68$  by 3 to turn the correction into scale units. The resulting value of  $-2.04$  represents the amount by which the assumed mean must be corrected.

<sup>2</sup> The question of how many places to carry decimals constantly arises in statistical work. For the computations here, carry the calculations to three decimal places and round them back to two. For example, the division of  $-25$  by 37 results in a decimal of  $-.675$ , which should be rounded back to  $-.68$ . If the decimal had been  $-.674$ , the decimal in the third position would have been dropped and the value would have become  $-.67$ .



(5) *Obtain the arithmetic mean.* The arithmetic mean results from the algebraic addition of the assumed mean and the correction. As the sign of the correction in the illustration is negative, the correction is subtracted from the assumed mean. Therefore, the arithmetic mean is  $51.00 - 2.04$ , or  $48.96$ . (*A.M.*) This step is equivalent in the illustration of Figure 25 to moving the fulcrum 2.04 score units downward to bring the beam into balance.

### Summary of steps in computing the arithmetic mean of grouped data

The steps below summarize the procedure outlined in detail above for computing the arithmetic mean (*A.M.*) from a grouped frequency distribution.

- (1) Assume a value for the mean. Choose the midpoint of an interval near the middle of the distribution.
- (2) Lay off the deviations from the assumed mean by intervals. Write in the *d* column a value of 0 for the interval in which the assumed mean lies and write values for other intervals by counting upward (positive signs) and downward (negative signs) by units.
- (3) Add a column at the right of the table and label it *fd*.
- (4) Obtain the correction to the assumed mean:
  - (a) Obtain the products of the frequencies and deviations by intervals. Multiply each frequency by its corresponding deviation, retaining negative signs for intervals below the assumed mean, and carry the results to the *fd* column,
  - (b) Determine the algebraic sum of the deviations of the scores. Algebraically add the values in the *fd* column obtained in step (4a) and retain the appropriate sign ( $\Sigma fd$ ),
  - (c) Determine the mean of the deviations of the scores. Divide the result of step (4b) by the number of cases, retaining the appropriate sign ( $\Sigma fd/N$ ), and
  - (d) Convert the mean of the deviations of the scores to the scale value. Multiply the result of step (4c) by the size of the class interval ( $c.i. \times \Sigma fd/N$ ).
- (5) Obtain the arithmetic mean. Algebraically add the correction of step (4d) to the assumed mean to obtain the arithmetic mean.  $A.M. = \text{Assumed Mean} + (c.i. \times \Sigma fd/N)$ .

### Exercises in Computing the Arithmetic Mean

4. Compute the arithmetic mean of the 40 arithmetic test scores tabulated in a frequency distribution for Problem 1, page 317. ( $A.M. = 54.90$ )
5. Compute the arithmetic mean of the 30 language test scores tabulated in a frequency distribution for Problem 2, page 317. ( $A.M. = 42.15$ )
6. Compute the arithmetic mean of the 30 spelling test scores tabulated in a frequency distribution for Problem 3, page 317. ( $A.M. = 13.03$ )

### Computing the mid-measure

Early workers with tests popularized the practice of taking the score of the middle paper of a pile of test papers arranged in order of size of scores as the expression of the central tendency of the group. The ease with which this so-called median is found has appealed to the classroom teacher. For a long time this was called the median. However, more recent workers with tests have recognized that the score of the middle paper in a pile of test papers stacked in order of size of scores is not the same as the middle point on the scale of a frequency table of these same scores. Accordingly, a distinction is now made between the score found on the middle paper of a pile of stacked papers and the median proper. The score of the middle paper of a pile of test papers arranged in systematic order is called the *mid-measure* to distinguish it from the *median*, which is the corresponding value when the data are grouped in a frequency distribution. Thus the mid-measure is a counting median found from ungrouped data. The median is computed only from tabulated data. The method of computing the mid-measure is illustrated by referring to the data given in Table 10, page 310, where these scores are arranged in descending order. *The mid-measure is the score of the middle paper when the number of cases is odd, or the average of the two scores nearest the middle when the number of cases is even.* In this case the number of papers or scores is 37 (odd). Thus the mid-measure is 48, a score such that there are just as many equal to or larger as there are equal to or smaller than it is.



## Computing the median of grouped data

By definition, the mid-measure and the median are quite similar, the main distinction being that the mid-measure is designated as an actual score on a certain paper (or the average of the scores on the two middle papers) while the median is defined directly in terms of a point on the scale of the frequency table on which it is based. *The median is a point on the scale such that 50 per cent of the cases in the distribution are above it and 50 per cent of the cases are below it.*

The method of computing the median from a grouped frequency distribution is presented below and illustrated in Table 15 for the same group of reading test scores used previously in this chapter.

(1) *Obtain the half-sum.* Divide the number of cases by two to determine how many of the cases fall below the median. For this illustration the half-sum, or  $N/2$ , is  $37 \div 2$ , or 18.5.

(2) *Obtain the sub-total.* Count upward into the distribution, adding the frequency for each successive interval, until exactly the half-sum or a number as closely approaching it as possible without exceeding it is reached. Thus, in the illustration,  $1 + 0 + 1 + 2 + 1 + 2 + 3 + 5 = 15$ . If the six scores in the interval 47-49 were added, the result, 21, would exceed the half-sum. The median, therefore, lies somewhere in the interval 47-49, for less than half of the scores lie below that interval and less than half of the scores lie above it.

(3) *Determine the correction.* The three steps involved are to determine the correction:

(a) *In terms of measures.* Subtract the sub-total from the half-sum. This subtraction will give the number of cases in the interval in which the median lies which must be added to the sub-total to obtain the half-sum, and consequently shows how much farther counting must continue upward to obtain the median. In the distribution of Table 15, this step becomes  $18.5 - 15 = 3.5$ .

(b) *In terms of the class interval.* Divide the result of the preceding step (half-sum — sub-total) by the number of cases in the interval in which the median falls. This will give the proportion of the interval that must be added to lower intervals in order to reach the point below which half of the cases fall. For the illustration of Table 15,  $3.5 \div 6 = .58$ . This step is based on the assumption that

the scores in an interval are uniformly distributed. More will be said about this assumption in a following section.

(c) *In terms of the scale distance.* Multiply the result of the preceding step by the size of the class interval so that the correction will be stated as a scale distance. Thus,  $.58 \times 3 = 1.74$  for the accompanying illustration.

TABLE 15. Computation of the median for the grouped frequency distribution of 37 reading test scores

Class Interval ( <i>c.i.</i> )		Frequency ( <i>f</i> )	
Integral Limits	Real Limits		
71-73	70.5-73.5	1	1. Obtain the half-sum $\frac{N}{2} = \frac{37}{2} = 18.5$
68-70	67.5-70.5	2	
65-67	64.5-67.5	1	2. Obtain the sub-total $1 + 0$ $+ 1 + 2 + 1 + 2 + 3 + 5$ $= 15$
62-64	61.5-64.5	1	
59-61	58.5-61.5	2	3a. Determine the correction (Measures) $18.5 - 15 = 3.5$
56-58	55.5-58.5	2	
53-55	52.5-55.5	3	3b. Determine the correction (Class interval) $3.5 \div 6 = .58$
50-52	49.5-52.5	4	
47-49	46.5-49.5	6	3c. Determine the correction (Scale distance) $.58 \times 3 = 1.74$
44-46	43.5-46.5	5	
41-43	40.5-43.5	3	4. Obtain the median $46.50 + 1.74 = 48.24$ ( <i>Mdn.</i> )
38-40	37.5-40.5	2	
35-37	34.5-37.5	1	
32-34	31.5-34.5	2	
29-31	28.5-31.5	1	
26-28	25.5-28.5	0	
23-25	22.5-25.5	1	
		$N = 37$	

(4) *Obtain the median.* Now add the correction in terms of scale distance to the lower real limit of the interval in which the median lies to obtain the median. The correction of 1.74 added to 49.50, or the lower real limit of the interval in Table 15 which contains the median, gives 48.24. (*Mdn.*)

Obviously, if the calculations of these steps were made by adding the frequencies down from the top of the distribution the median



would be the same. In that case, 16 scores falling above the interval 47-49, the correction of 1.26 (2.5 measures, .42 of an interval, 1.26 in terms of scale units) would be subtracted from 52.5, the top of the step, to give the same result of 48.24 for the median.

### Summary of steps in computing the median

The steps listed below provide in form for easy use the procedures necessary in computing the median from a grouped frequency distribution.

- (1) Obtain the half-sum. Divide the number of cases by 2. ( $N/2$ )
- (2) Obtain the sub-total. Count upward into the distribution by adding successive frequencies until a number exactly equal to the half-sum or as closely approaching it as possible without exceeding it is reached.<sup>3</sup>
- (3) Determine the correction:
  - (a) In terms of measures by subtracting the sub-total from the half-sum,
  - (b) In terms of the class interval by dividing the result of step (3a) by the number of cases in the interval in which the median lies, and
  - (c) In terms of the scale distance by multiplying the result of step (3b) by the size of the class interval.
- (4) Obtain the median. Add the correction of step (3c) to the lower real limit of the interval in which the median lies to obtain the median. (*Mdn.*)

### Exercises in Computing the Mid-Measure and Median

7. Find the mid-measure of the 40 arithmetic test scores of Problem 1, page 317.
8. Find the mid-measure of the 30 language test scores of Problem 2, page 317.
9. Compute the median of the 40 arithmetic test scores tabulated in a frequency distribution for Problem 1, page 317. (*Mdn.* = 53.79)
10. Compute the median of the 30 language test scores tabulated in a frequency distribution for Problem 2, page 317. (*Mdn.* = 43.50)

<sup>3</sup> If exactly the half-sum is reached, the median is usually the upper real limit of the interval the frequency of which was last added in the counting process. However, if the next higher interval should happen to have a zero frequency, the median is the midpoint of that interval.

### Assumptions in computing measures of central tendency

As has been indicated briefly in a preceding section of this chapter, the assumption concerning the distribution of scores within each class interval varies according to which of the measures of central tendency is being computed. Figure 26 shows in parallel graphical representations of the distribution of scores assumed in the computation of the arithmetic mean and the median for several class intervals near the center of the distribution used for illustrative purposes in the preceding pages.

<i>c.i.</i>	<i>f</i>	Arithmetic Mean		Median	
		Real Limits	Midpoints	Real Limits	Divisions between Measures
56-58	2	58.5 —————	57.0	58.5 —————	58.50
				2	57.00
				1	
53-55	3	55.5 —————	54.0	55.5 —————	55.50
				3	54.50
				2	53.50
50-52	4	52.5 —————	51.0	52.5 —————	52.50
				4	51.75
				3	51.00
47-49	6	49.5 —————	48.0	2	50.25
				1	49.50
				6	49.00
44-46	5	46.5 —————	45.0	5	48.50
				4	48.00
				3	47.50
		43.5 —————		2	47.00
				1	46.50
				5	45.90
				4	45.30
				3	44.70
				2	44.10
				1	43.50

Fig. 26. Assumptions concerning the distribution of scores in class intervals in the computation of the arithmetic mean and median

In the computation of the arithmetic mean it is assumed that each score in a grouped frequency distribution has the value of the mid-point of the interval in which it is tabulated. This is illustrated by



the heavily ruled lines in the left-hand portion of Figure 26. On the other hand, in the computation of the median it is assumed that each score in a grouped frequency distribution expands or contracts in such manner that it shares the scale distance through a class interval equally with the other measures in the same class interval. This assumption is illustrated in the right-hand portion of Figure 26. Thus, each of the five measures in the interval 44-46 is assumed to have the value of 45 in computing the arithmetic mean and to occupy one-fifth of the scale distance through that interval ( $\frac{1}{5} \times 3 = 0.6$ ) in computing the median. Again, the three scores in the step 53-55 are assumed in computing the arithmetic mean to be concentrated at 54, the midpoint of the interval, while in computing the median each of the scores is assumed to occupy one-third of the scale distance through that interval ( $\frac{1}{3} \times 3 = 1.0$ ).

This leads to one further important distinction between the arithmetic mean and the median. The mean is algebraic in nature (although the various operations can be stated either in algebraic or in arithmetic terms), while the median is arithmetic in nature.

### 3 MEASURES OF VARIABILITY

#### Need for measures of variability

The measures of central tendency represented by the arithmetic mean and the median are valuable statistical measures but they describe only one characteristic of the data—the tendency of the scores to pile up at or near the middle of the distribution. Descriptions of test results based wholly on one or the other of these measures are incomplete.

The two groups of scores presented as Class A and Class B in Table 16 illustrate this situation clearly. The means of the two series of scores are identical, each being 86. The *range* of the scores for Class A is 72 ( $122 - 50$ ), which is exactly three times the range ( $98 - 74 = 24$ ) of scores for Class B. Even an inexperienced teacher will recognize that very different ranges of ability are present in these two classes and that correspondingly different instructional problems are presented to the teacher.

A graphical illustration based on other data showing the unlikenesses that may appear in distributions having the same mean is given in Figure 27.

TABLE 16. Data showing identical means but unlike variability

Class A		Class B
122		98
116		96
108		95
101		93
96		90
92		89
89		87
86	<i>A.M.</i>	86
83		85
80		83
76		82
71		79
64		77
56		76
50		74

It is a common practice to show frequency distributions in graphical form by representing the frequencies at a given point on the scale in terms of a line erected perpendicular to the base line or scale. If

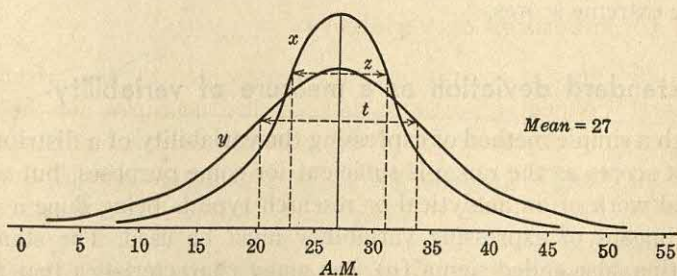


Fig. 27. Measures of variability for homogeneous and heterogeneous data

the tops of a large number of these perpendiculars are connected, the result is a curved line that usually is close to the base line at the ends of the scale but that rises quite rapidly from the base line near the middle of the scale. In Figure 27 the curve marked *x* represents the distribution of scores made on a certain test by a class. The closeness of the curve to the base line at the ends or extremes shows that there are relatively few very low and very high scores. The high



point of the curve near the middle indicates that a great many pupils made scores near the average. This is typical of situations usually found where considerable numbers of cases are involved.

It will be noted in Figure 27 that, while the middle portion of the curve  $x$  rises much higher than the similar portion of the curve  $y$ , the extremes of curve  $x$  do not go out on the base line in either direction so far as is true of curve  $y$ . This flatness or peakedness of the curve is the graphical indication of the variability of the data it represents. The less peaked the curve the greater the variability, other things being equal. It is thus apparent from this illustration that while the means of these two distributions are identical, very greatly different teaching problems are represented. Curve  $x$  represents a relatively homogeneous group, while curve  $y$  represents a more widely scattered group.

### The range as an expression of variability

The *range* of scores, that is, the scale distance between the lowest and highest scores in a distribution, is one way of expressing variability. However, it is one of the least reliable measures of variability or dispersion, since it is apparent that it is affected by the fluctuation of the extreme scores.

### The standard deviation as a measure of variability

Such a simple method of expressing the variability of a distribution of test scores as the range is sufficient for some purposes, but where careful work of an analytical or research type is being done a more exact means of expressing variability must be used. The standard deviation, also called sigma ( $\sigma$ ), has many characteristics that make it a useful measure of variability. *The standard deviation is a sort of arithmetic mean of the squares of the deviations from the mean of the distribution.* It is a special type of mean of the deviations because of the method used in computing it. In calculating the standard deviation ( $\sigma$ ), each individual deviation from the mean is squared, the sum of these squared values is then divided by the number of such deviations, and the square root of the result is then obtained. Restated, *the standard deviation (sigma) is the square root of the mean of the squares of the deviations from the mean of a distribution.* Ex-

pressed in symbols the standard deviation becomes *c.i.*  $\sqrt{\frac{\Sigma fd^2}{N} - c^2}$ , where  $\Sigma fd^2$  equals the deviations in the form of the sum of the products of the frequencies in each interval by the deviation of each interval from the assumed mean,  $N$  equals the number of cases in the distribution,  $c$  equals the correction as found in calculating the mean, and *c.i.* equals the size of the class interval of the distribution in units.

The lines  $S$  and  $T$  of Figure 28 represent ordinates erected at a distance equal to one  $\sigma$  on either side of the mean. The standard deviation takes into account approximately 68 per cent (in a normal distribution 68.26%) of the area of such a distribution. That is, ordinates erected at a distance equal to one sigma on either side of the arithmetic mean include approximately two-thirds of the cases in the distribution.

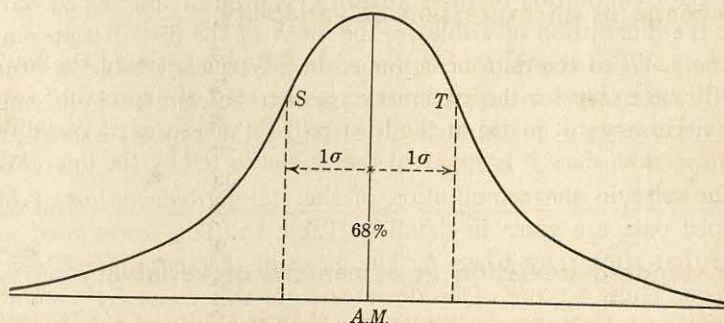


Fig. 28. Arithmetic mean and standard deviation

A further interesting characteristic of the standard deviation is also indicated in Figure 28. Mathematically, the value *sigma* bears a definite relationship to the curve of the distribution itself. Where any large number of cases or scores are found in a distribution there is a tendency for the larger portion of the cases to pile up at or near the middle of the distribution. When a normal distribution is presented in graphical form, the result is a symmetrical bell-shaped curve with many cases in the middle and few cases at the extremes. Certain types of these characteristic bell-shaped distributions have come to be called *normal curves*. For these normal curves, formulae



have been derived from which such typical curves may be constructed if certain basic data concerning the curve are given. In these formulae *sigma* is one of the values that must be given in order to construct such a curve. Sigma, in the typical formula, represents the distance from the mean at which the curve changes from convex to concave. In Figure 28 the points where the curve changes its character are indicated by the ordinates lettered *S* and *T*.

Thus, because of this direct mathematical relationship that the standard deviation bears to the curve of the distribution itself and the reliable expression of variability that it provides, since every deviation in the distribution is considered, the standard deviation is one of the most useful of the measures of variability.

### Computing the standard deviation of ungrouped data

In the computation of the standard deviation of ungrouped data, as in the illustration of Table 17, the mean of the distribution must be found. When the data are grouped in a frequency table it is not strictly necessary for the arithmetic mean to be computed, although it is necessary to go through all steps of the process except the last.

The steps in the computation of the standard deviation of ungrouped data are given in detail in Table 17. The scores used are those that appear for Class A. The mean of the scores for Class A is 86.00. Thus, a score of 89 deviates from this mean by 3 points. A score of 96 deviates 10 points. Other deviations are similarly shown in the *d* column of Table 17. The standard deviation ( $\sigma$ ) is the square root of the mean of the squares of these deviations from the mean of the scores. It is necessary, therefore, to square each of these deviations. These squares are given under the column headed  $d^2$ . Since each deviation appears only once and the data are ungrouped, the formula may be simplified to read  $\sigma = \sqrt{\frac{\sum d^2}{N}}$ . The sum of the deviations squared ( $\sum d^2$ ) in this case is 6100. The mean of these squared deviations is therefore 406.67. Therefore, to turn it into units of the scale, the square root of this quantity must be taken. This value is 20.17, which is the standard deviation ( $\sigma$ ) of this series of scores. The mean of this distribution is 86.00 and the  $\sigma$  is 20.17. Hence, approximately two-thirds of the scores will be found between scores 20.17 points larger and 20.17 points smaller than this mean.

TABLE 17. Computation of the standard deviation for ungrouped data

Test Scores	Deviations ( <i>d</i> )	Deviations Squared ( <i>d</i> <sup>2</sup> )	Computations
122	+36	1296	$\sigma = \sqrt{\frac{\sum d^2}{N}}$ $= \sqrt{\frac{6100}{15}}$ $= \sqrt{406.67}$ $= 20.17$ $A.M. = \frac{1290}{15}$ $= 86.00$
116	+30	900	
108	+22	484	
101	+15	225	
96	+10	100	
92	+ 6	36	
89	+ 3	9	
86	0	0	
83	- 3	9	
80	- 6	36	
76	-10	100	
71	-15	225	
64	-22	484	
56	-30	900	
50	-36	1296	
1290		$\sum d^2 = 6100$	

### Computing the standard deviation of grouped data

The method of computing the standard deviation from ungrouped data illustrated in Table 17 may be applied with few changes to the calculation of sigma from grouped data. A slight change in the general formula is required, for when the scores are grouped in class intervals the deviations of the scores must be considered by groups having the midpoints of the intervals in which they are found. This permits the expression of the deviations in class intervals rather than in units of the scale. The formula for use in calculating the standard deviation when the data are grouped in a frequency distribution is

$$c.i. \sqrt{\frac{\sum fd^2}{N} - c^2}.$$

The steps in the application of this formula in the calculation of the standard deviation of the scores originally presented in Table 9 will make clear all of the processes involved. The computations them-



selves are shown in Table 18. The first five steps of procedure for computing the standard deviation are identical with those given above for determining the arithmetic mean.

(1) *Assume a value for the mean.* Assume a mean as near as possible to the arithmetic mean of the distribution in order that the correction ( $c$ ) may be as small as possible. In Table 18, as for the computation of the *A.M.* for the same distribution, the assumed mean is taken as the midpoint of the interval 50-52.

TABLE 18. Computation of the standard deviation for the grouped frequency distribution of 37 reading test scores

<i>c.i.</i>	<i>f</i>	<i>d</i>	<i>fd</i>	<i>fd</i> <sup>2</sup>	
71-73	1	+7	+ 7	49	1. Assume a value for the mean (51.00)
68-70	2	+6	+12	72	2. Lay off the deviations from the assumed mean by intervals
65-67	1	+5	+ 5	25	3. Obtain the correction to the assumed mean $\Sigma fd = -25$ ; $\Sigma fd/N = -25/37 = -.676$ ( $c$ )
62-64	1	+4	+ 4	16	4. Square the correction to the assumed mean $c^2 = -.676^2 = .457$ ( $c^2$ )
59-61	2	+3	+ 6	18	5. Add the $fd^2$ column to the table
56-58	2	+2	+ 4	8	6a. Obtain the squared deviations of the scores ( $fd^2$ )
53-55	3	+1	+ 3	3	6b. Obtain the sum of the squared deviations of the scores $\Sigma fd^2 = 503$
50-52	4	0	0	0	6c. Obtain the mean of the squared deviations of the scores $\Sigma fd^2/N = 503/37 = 13.595$
47-49	6	-1	- 6	6	6d. Obtain the corrected mean of the squared deviations of the scores
44-46	5	-2	-10	20	$\frac{\Sigma fd^2}{N} - c^2 = 13.595 - .457 = 13.138$
41-43	3	-3	- 9	27	7a. Obtain the standard deviation (Class intervals)
38-40	2	-4	- 8	32	$\sqrt{\frac{\Sigma fd^2}{N} - c^2} = \sqrt{13.138} = 3.62$
35-37	1	-5	- 5	25	7b. Obtain the standard deviation (Scale units)
32-34	2	-6	-12	72	$c.i. \sqrt{\frac{\Sigma fd^2}{N} - c^2}$
29-31	1	-7	- 7	49	$= 3 \times 3.62 = 10.86$ (S.D.)
26-28	0	-8	0	0	
23-25	1	-9	- 9	81	
<i>N</i>	37				

(2) *Lay off the deviations from the assumed mean by intervals.* Lay off deviations above and below the interval in which the mean is assumed to lie. Signs of deviations below the assumed mean are negative.

(3) *Obtain the correction to the assumed mean.* Multiply the frequency in each interval by its corresponding deviation and carry the results to the  $fd$  column. Take account of signs. Obtain the sum of the positive  $fd$  and the negative  $fd$  values separately and then determine the algebraic sum of the entire column. Take proper account of the signs of the two values first obtained. In the illustration of Table 18, the  $\Sigma fd$  is  $-25$ . Then obtain  $\Sigma fd/N$  to determine the correction ( $c$ ). For the distribution of the table, the correction is  $-25/37$  or  $-.676$ .<sup>4</sup> In this computation the correction is left in terms of class intervals and is not, as is the case in computing the arithmetic mean, converted into scale units.

(4) *Square the correction to the assumed mean.* Square the correction in class-interval units<sup>5</sup> to obtain the second term under the radical sign ( $c^2$ ) in the formula given above for the standard deviation. For the illustration this becomes  $-.676^2$ , or  $.457$ .

(5) *Add a column at the right of the table and label it  $fd^2$ .* This column is used for recording the squares of the deviations of the scores from the assumed mean.

(6) *Obtain the mean of the squared deviations of the scores.* Obtain this mean by use of the following steps of procedure:

(a) *Obtain the squared deviations of the scores.* Multiply each value in the  $fd$  column by its corresponding  $d$  value and place the results in the  $fd^2$  column. All signs will be positive.

(b) *Obtain the sum of the squared deviations of the scores.* Add the values in the  $fd^2$  column to obtain  $\Sigma fd^2$ , or the sum of the squared deviations of the scores from the assumed mean. The  $\Sigma fd^2$  of Table 18 is 503.

(c) *Obtain the mean of the squared deviations of the scores.* Divide  $\Sigma fd^2$  by  $N$  to obtain the mean of the squared deviations of the scores from the assumed mean, or  $\Sigma fd^2/N$ . This value for the data of Table 18 is  $503/37$ , or  $13.595$ .

(d) *Obtain the corrected mean of the squared deviations of*

<sup>4</sup> Computations for the standard deviation should uniformly be carried to four decimal places and rounded back to three decimal places, otherwise using the same procedure for rounding numbers as was illustrated on page 321. The standard deviation itself is commonly stated to two decimal places only, however.

<sup>5</sup> The student will find the tables of squares and square roots such as are given in Lindquist, *A First Course in Statistics*, Revised edition (Houghton Mifflin Co., Boston, 1942), pages 230-39, very helpful in his work from this point on. Use of such tables, a slide rule, or an electrical calculating machine should speed up his work and result in a high degree of accuracy.



the scores. Subtract the square of the correction ( $c^2$ ) from  $\Sigma fd^2/N$  to obtain the corrected mean of the squared deviations of the scores from the arithmetic mean, i.e., to account for the deviation of the assumed mean from the arithmetic mean.<sup>6</sup> The result for the accompanying illustration is  $13.595 - .457$ , or  $13.138$ .

(7) *Obtain the standard deviation.* Complete the process of finding the standard deviation by using the following two steps of procedure:

(a) *In terms of class intervals.* To obtain the standard deviation in units of class interval, extract the square root of the mean of the squared deviations of the scores from the arithmetic mean, i.e., obtain  $\sqrt{\frac{\Sigma fd^2}{N} - c^2}$ . The square root of  $13.138$  in the accompanying illustration is, to two decimal places,  $3.62$ .

(b) *In terms of the scale distance.* To put the standard deviation into scale units, multiply the square root of the mean of the squared deviations of the scores from the arithmetic mean by the size of the class interval, i.e., obtain  $c.i. \sqrt{\frac{\Sigma fd^2}{N} - c^2}$ . For the data of Table 18, this becomes  $3 \times 3.62$ , or  $10.86$ .

This means that approximately 68 per cent of the scores will be found between ordinates erected at a distance of  $10.86$  score units on either side of the mean. Of course this will not be strictly true, since no distribution having as few scores as 37 is likely to approximate the normal curve very closely. However, approximately two-thirds of the scores may be expected to lie between the two points one standard deviation above and one standard deviation below the arithmetic mean. For this illustration the points are  $48.96 + 10.86$  and  $48.96 - 10.86$ , or  $59.82$  and  $38.10$ . Actually 25 scores, or 67.6 per cent of the total, lie between these two points.

### Summary of steps in computing the standard deviation of grouped data

The steps of procedure for computing the standard deviation from a grouped frequency distribution are as follows:

<sup>6</sup> The square of the correction ( $c^2$ ) is always subtracted from  $\Sigma fd^2/N$  because the latter value is always too large in instances where the assumed mean and arithmetic mean are not identical. If  $c$  has any value other than zero, the assumed mean and the arithmetic mean do not coincide and the deviations computed about the assumed mean are too large. The square of the correction ( $c^2$ ) must be subtracted from  $\Sigma fd^2/N$  to compensate for this difference.

- (1) Assume a value for the mean. Follow the procedure of step (1), page 322, for computing the arithmetic mean.
- (2) Lay off the deviations from the assumed mean by intervals. Follow the procedure of step (2), page 322, for computing the arithmetic mean.
- (3) Obtain the correction to the assumed mean. Follow the procedures of steps (4a), (4b), and (4c), page 322, for computing the arithmetic mean. As the correction ( $c$ ) is to be stated in class-interval rather than in scale units, do not include step (4d).
- (4) Square the correction to the assumed mean. Square the result of step (3) to obtain  $c^2$ .
- (5) Add a column at the right of the table and label it  $fd^2$ .
- (6) Obtain the mean of the squared deviations of the scores.
  - (a) Obtain the squared deviations of the scores. Multiply each  $fd$  value by its corresponding  $d$  value and write the products in the  $fd^2$  column.
  - (b) Obtain the sum of the squared deviations of the scores. Add the values in the  $fd^2$  column to obtain  $\Sigma fd^2$ .
  - (c) Obtain the mean of the squared deviations of the scores. Divide the result of step (6b) by the number of cases to obtain  $\Sigma fd^2/N$ .
  - (d) Obtain the corrected mean of the squared deviations of the scores. Subtract the result of step (4) from the result of step (6c) to obtain  $\frac{\Sigma fd^2}{N} - c^2$ .
- (7) Obtain the standard deviation:
  - (a) In terms of class intervals. Extract the square root of the result of step (6d) to obtain  $\sqrt{\frac{\Sigma fd^2}{N} - c^2}$ .
  - (b) In terms of the scale distance. Multiply the result of step (7a) by the size of the class interval to obtain the standard deviation in scale units.  $S.D. = c.i. \sqrt{\frac{\Sigma fd^2}{N} - c^2}$ .

## Exercises in Computing the Standard Deviation

11. Compute the standard deviation of the 40 arithmetic test scores tabulated in a frequency distribution for Problem 1, page 317.  
( $S.D. = 11.19$ )
12. Compute the standard deviation of the 30 language test scores tabulated in a frequency distribution for Problem 2, page 317.  
( $S.D. = 18.00$ )



## Selected References

- FROEHLICH, CLIFFORD P., AND DARLEY, JOHN G. *Studying Students: Guidance Methods of Individual Analysis*. Chicago: Science Research Associates, 1952. Chapter 2.
- GARRETT, HENRY E. *Statistics in Psychology and Education*. Fourth edition. New York: Longmans, Green and Co., 1953. p. 1-9, 20-24; Chapters 2-3.
- GREENE, EDWARD B. *Measurements of Human Behavior*. Revised edition. New York: Odyssey Press, 1952. p. 363-70.
- GREENE, HARRY A., AND CRAWFORD, JOHN R. *Work-Book in Educational Measurements and Evaluation*. New York: Longmans, Green and Co., 1945. Units 3-6.
- LEE, J. MURRAY. *A Guide to Measurement in Secondary Schools*. New York: D. Appleton-Century Co., Inc., 1936. p. 436-52.
- LINDQUIST, E. F. *A First Course in Statistics*. Revised edition. Boston: Houghton Mifflin Co., 1942. Chapters 1-2, 5-6.
- MICHEELS, WILLIAM J., AND KARNES, M. RAY. *Measuring Educational Achievement*. New York: McGraw-Hill Book Co., Inc., 1950. p. 423-36, 439-42.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. p. 83-94.
- ODELL, C. W. *An Introduction to Educational Statistics*. New York: Prentice-Hall, Inc., 1946. Chapters 2, 4, 6.
- REMMERS, H. H., AND GAGE, N. L. *Educational Measurement and Evaluation*. New York: Harper and Brothers, 1943. p. 512-26.
- ROSS, C. C. *Measurement in Today's Schools*. Second edition. New York: Prentice-Hall, Inc., 1947. p. 218-37.
- WALKER, HELEN M. *Elementary Statistical Methods*. New York: Henry Holt and Co., 1943. p. 28-36; Chapters 7-8.
- WALKER, HELEN M. "Statistical Understandings Every Teacher Needs." *Teachers College Record*, 49:452-57; April 1948.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. p. 43-59.
- WEITZMAN, ELLIS, AND MCNAMARA, WALTER J. *Constructing Classroom Examinations*. Chicago: Science Research Associates, 1949. p. 112-28.

## ***Interpreting the Results of Measurement***

THIS CHAPTER gives consideration to the following points concerning the interpretation of results from measurement:

- A. Meaning of test scores.
- B. Formal and informal types of derived scores.
- C. Graphical representations—frequency polygons and histograms.
- D. Cumulative frequency graphs in estimating percentiles and percentile ranks.
- E. Norms—age, grade, and percentile types.

The major techniques used in summarizing and describing single sets of test scores were presented in Chapter 12. Various techniques and devices for attributing additional meaning to test scores and distributions of test results are necessary, however, if optimum use of such results is to be attained. This chapter deals with the basic procedures and devices commonly used in attributing readily understandable meaning to the results of measurement.

### **1 TEST SCORES**

The problems of summarizing test scores and of interpreting the results revealed by these summaries are very closely related and, were it not for the length and detail of the discussion they require, would probably be considered in a single chapter. The preceding chapter deals with three of the six major problems of statistical pro-



cedure as related to test analysis. This chapter concerns itself almost entirely with two more of these problems of test interpretation. The last problem will be treated in Chapter 14.

## Meaning of a test score

Test scores are valuable to the classroom teacher to the extent that they can be interpreted. It is therefore important to define clearly what is meant by a test score. To do this, two or three new concepts require explanation. In the first place, *a test score is a numerical expression of performance* on the part of an individual. Sometimes the test score is merely the number of items answered correctly. Again it may be an arbitrarily defined scale value. But whatever its form, its function is to reveal in a quantitative way the performance of an individual as he responds to stimuli given under certain conditions.

This leads to the second concept involved in the meaning of a score. The test score is an evidence of performance. Performance, the response of the individual to the test situation, is taken to mean in educational measurements the expression of ability operating under certain conditions. *Performance may be thought of as Ability + Conditions*. Scores on tests are definitely influenced by conditions. The pupil may make a low score because he does not have the ability to do better. On the other hand, he may make a low score because of illness, discomfort, poor illumination, a broken pencil, indifference for the subject, dislike for the teacher or examiner, failure to give attention to and to comprehend the directions, or any one of many other reasons. Accordingly, there is the possibility and even the likelihood of a serious error in the assumption that a test score is direct evidence of ability. The conditions under which the performance takes place must be known before it is safe to infer ability from performance.

Ability, as an abstract concept, may be defined as the power to do. Power to do, to respond to stimuli and to situations, is the product of training and experience. *Ability may be thought of as Capacity + Training*, which suggests that unless training and native capacity factors are known inferences about abilities may be misleading. This point becomes particularly serious in the interpretation of intelligence test results, for it is a common practice for users of such tests to infer capacity (mental ability) from performance scores. The real

seriousness of this type of uncritical inference may be seen by comparing the interpretations of an achievement test score and an intelligence test score. Both are basically expressions of performance. Equal abilities may be inferred from equal scores in both types of tests if and when the conditions under which they are given are all definitely under control. While it is difficult to make sure that all physical and psychological factors are adequately controlled in a testing situation, it is possible to regulate most of the mechanical conditions within reasonable limits. The significant point to note here, however, is the fact that users of achievement tests stop with an inference of equality of ability from equal performance scores, but users of intelligence tests are obliged to make a further inference.

In the interpretation of intelligence test results, it is common practice to infer equal capacity from apparent evidences of equal abilities. The fallacies in this argument and the dangers of this step must be readily apparent. Equal capacities may be inferred from performance scores only when there is direct and positive evidence of two things: first, that the conditions under which the testing took place were identical and equally well controlled; second, that the training opportunities of the individuals compared have been equal. The mechanics of testing now make it fairly easy to control testing conditions. The second factor represents a real stumbling block in the way of an accurate and sane interpretation of intelligence test results. The naïve manner in which some makers and many users of such tests assume equality of learning opportunity, and hence equal capacity from equal performance scores, is one of the things that has made many teachers and students skeptical of their value.

The foregoing discussion of the meaning of a test score may appear to indicate that it is impossible to give meaning to any kind of test score. Such is not the intention, even though the purpose here is to emphasize the need for a conservative attitude in test score interpretation. In the long run, the more that is known about the variables underlying test scores, the more critical must the user become. The greatest damage that has been done to the field of educational measurements has come as a direct result of carelessness and ignorance on the part of users of tests, and their tendency to draw unwarranted conclusions from the results. The teacher should be able critically to select suitable tests and scales for classroom use, to control the mechanical conditions of their administration, and to draw sane and defensible conclusions and inferences from the results.



## Giving meaning to test scores

The user of educational tests in the classroom is confronted with two types of test data for interpretation. The first type, and undoubtedly the more common of the two, deals with the results of informal, teacher-made tests. The results from these classroom tests are in turn of two types—the subjective scores assigned by teachers to pupils' responses to essay tests and the scores resulting from informal objective examinations. While something can be done to improve the interpretation of the relatively unreliable marks assigned to the discussion-type exercise, much more is possible in the accurate interpretation of the scores resulting from the use of objective examinations. Since one of the major functions of the standardization of a test is the establishment of meaning for the test scores, many additional types of interpretation are thus made possible in the use of the second type of test data, which is obtained from the use of standardized tests.

## 2 FORMAL TYPES OF DERIVED SCORES

### Function of derived scores

Test scores used to describe the performance of pupils are expressed in a variety of different units and in relation to a variety of different scales of measurement. In some tests the unit of measurement used may be relatively large. Pupil scores expressed in terms of these large units are often numerically small. A long test composed of many items may result in numerically large scores. It is thus apparent that a given score on one test may represent an exceptionally good performance while the same score on another test may represent an exceedingly poor performance. Some common basis must be established if comparisons of scores based on these widely different types of scales are to be possible. A number of methods have been developed for the calculation of derived scores which will partially take care of this difficulty.

### Relation of derived scores to norms

Confusion may easily exist in the thinking of the student concerning the distinction between derived scores and norms. As a matter of fact, tables of norms often yield such derived scores as grade scores,

age scores, or percentile ranks. The use of norm tables for obtaining such derived scores directly from raw scores or point scores on various tests is illustrated in Chapter 5 and is also discussed later in this chapter. Such ratios as the intelligence quotient, educational quotient, accomplishment quotient, and reading quotient are derived scores, but they are obtained by a division of one value by another. Some of these quotients are presented later in this chapter and others are treated in Chapter 10.

Another possible source of confusion to the student lies in the fact that some tests provide a two-step procedure from raw scores to norms. In those cases, such derived scores as standard scores, scaled scores, converted scores, and equated scores are obtained from raw scores. Such derived scores have more meaning than do raw scores, but they do not always have final meaning for the interpretation of test results. Consequently, it is often necessary to enter a table of norms with the derived scores and to interpret them in terms of such other derived scores as grade scores, age scores, or percentile ranks. Situations of this type will also be discussed in a later section of this chapter.

It is believed that the most satisfactory method of familiarizing the student with derived scores and norms is first to present the various types of derived scores, methods of computing them, and something of their meaning, and then to illustrate the need for and use of norms in the further interpretations required to make some of them meaningful. In the treatment that follows, three types of derived scores are distinguished: (1) those based on average or median performance, (2) quotients and related measures, and (3) those based on variability of performance.

### Derived scores based on average performance

The two types of derived scores based on average or median performance are the grade score and the age score. These are directly dependent upon tables of norms, for it is only by entering norm tables with raw scores or some other forms of scores that grade equivalents or age equivalents can be determined. The meaning of grade scores and age scores is presented here and the use of tables of norms for their derivation is illustrated later in this chapter.

*Grade equivalents.* A grade equivalent indicates the position on a grade scale at which a pupil's test performance places him. For



example, a child may attain a score on a reading test that is identical with the average or median score of pupils three months into the fourth grade. If so, his grade equivalent on the subject matter of the test is 4.3, regardless of whether he may be in the fourth grade or in some grade above or below the fourth. Grade scores are sometimes referred to as *G*-scores or, less frequently, as *B*-scores. Grade and months are commonly listed as a number and decimal respectively or as a number and exponent respectively. Thus the above grade equivalent might be stated either as 4.3 or as  $4^3$ .

*Age equivalents.* In the manner very similar to that which operates for grade scores, age equivalents indicate the position on an age scale at which a pupil's test performance places him. The hypothetical child whose reading test score gave him a grade equivalent of 4.3, for example, might be found by the use of a table of age norms to have an age equivalent of nine years eight months (9-8) on the same test. This would mean that his score was identical with the score made by the average or median pupil nine years and eight months of age. He might actually be a year or so older or younger; his age equivalent on the subject matter of the test would nevertheless be 9-8. Age equivalents are represented by such terms as educational age (*EA*) for achievement over broad areas of subject matter, mental age (*MA*) for performance on general intelligence tests, and reading age (*RA*) for achievement in reading. Such ages are commonly stated in hyphenated form, the first number indicating years and the second number months of age. Thus the *EA* of 9-8 indicates that in broad educational achievement the child used in the above illustration is at the same level as average children nine years and eight months of age.

Although this book is most directly concerned with the types of age equivalents noted above, the same technique is applied to the measurement of other aspects of child growth and performance. For example, anatomical age, physiological age, and social age are comparable terms that are employed with varying degrees of exactness in meaning. Chronological or life age is, of course, the most widely used of all, and is frequently employed as the basic or criterion measure of test validity, as will be pointed out in the following paragraphs.

## Quotients as derived scores

Quotients and other similar derived scores show the relationship existing between two characteristics for the child as a means of indicating the manner in which growth of various types is related. For instance, the educational quotient, intelligence quotient, and reading quotient are ratios respectively between a child's educational or mental and chronological ages. The accomplishment quotient is the ratio between a child's educational and mental ages. The first three are based on the idea that on the average a child grows in all ways more nearly in conformance with his chronological age than with any other measures, and also upon the recognition that deviations from that pattern of growth result from individual differences and are meaningful in the guidance of the child. The accomplishment or achievement quotient is based on the idea that the child's mental age is perhaps a better criterion by which to judge his educational growth than is his chronological age. All of these have been discussed in appropriate chapters elsewhere in this volume.

Computation of the various quotients listed above will be illustrated for a pupil who has a chronological age (*CA*) of 8-4, an educational age (*EA*) of 9-2, a mental age (*MA*) of 9-7, and a reading age (*RA*) of 9-4. The last three ages would be determined in the manner indicated in the above section from his scores on a general achievement, a general intelligence, and a reading test. The quotients are all based on computations in which each age is reduced to months, and all ratios are multiplied by 100 to eliminate the use of decimals in the results.

For the child whose various age levels or age equivalents are given above, his educational quotient (*EQ*) would be

$$EQ = 100 \frac{EA}{CA} = 100 \frac{9-2}{8-4} = 100 \frac{110 \text{ (months)}}{100 \text{ (months)}} = 110,$$

his intelligence quotient (*IQ*) would be

$$IQ = 100 \frac{MA}{CA} = 100 \frac{9-7}{8-4} = 100 \frac{115 \text{ (months)}}{100 \text{ (months)}} = 115,$$

his reading quotient (*RQ*) would be

$$RQ = 100 \frac{RA}{CA} = 100 \frac{9-4}{8-4} = 100 \frac{112 \text{ (months)}}{100 \text{ (months)}} = 112,$$



and his accomplishment quotient (AQ) would be

$$AQ = 100 \frac{EQ}{IQ} = 100 \frac{110}{115} = 95.6 \text{ or } 96,$$

or

$$AQ = 100 \frac{EA}{MA} = 100 \frac{9.2}{9.7} = 100 \frac{110}{115} = 95.6 \text{ or } 96.$$

These quotients indicate that the child is well above average for his age in intelligence and is somewhat less accelerated educationally. Within the limits of reliability for the *AQ*, discussed in some detail in Chapter 10, it appears that his achievement is not quite what might be expected of a child of his mental ability level. His reading quotient indicates a somewhat greater advancement in that subject than for the average of all other areas of achievement covered by the general achievement test from which his *EA* was determined.

With this brief presentation of the method of deriving the various commonly used quotients as a background, the student should be able to interpret these quotients adequately when he encounters them elsewhere in this volume. It should be understood that the *RQ* is merely representative of quotients that can be derived for the various subjects of the curriculum if age norms are given for such subjects on the standardized tests used. In practice, such quotients are seldom used except for reading and arithmetic, however.

### Derived scores based on variability of performance

Derived scores based on variability of performance are of two types: (1) percentile ranks, and (2) scores that express position on a scale in units of the standard deviation or quartile deviation. Although these methods are similar in some respects, they differ in several fundamental ways which determine their effectiveness for certain types of uses.

Percentile ranks are less reliable than are derived scores based on the standard deviation because they are more affected by minor irregularities in the distribution of scores upon which they are based than is the standard deviation. Percentile ranks cannot with strict validity be averaged, whereas averaging several scores similarly stated in terms of the *S.D.* is a defensible procedure. Measures of rank are based on equivalent areas under the curve, so that percentile ranks near the middle of the distribution, for example of

48 and 49, usually represent closely similar raw scores, whereas percentile ranks near an extreme of the distribution, say of 2 and 3, may well represent raw scores differing by a number of points. On the other hand, derived scores based on the *S.D.* differ by equivalent distances along the scale, so that they represent merely the application of a new and more meaningful linear scale to a linear distance originally represented in raw-score units.

*Percentile ranks.* The test performance of a pupil may be expressed in terms of his position in the distribution of scores for pupils in his school grade, in a certain course he is taking, such as plane geometry, or who, with him, have studied a certain subject, such as a foreign language, for a given number of semesters. This is accomplished by dividing the distribution so that the various divisions contain the same percentage of the total number of cases. Various plans are to divide the distribution into fourths, fifths, tenths, or hundredths. The distribution is divided into fourths by computing the quartiles ( $Q_1$ , median, and  $Q_3$ ); into fifths by computing the quintiles, i.e., percentiles 20, 40, 60, and 80; into tenths by computing the deciles, i.e., percentiles 10, 20, 30, . . . 70, 80, and 90; and into hundredths (percentile ranks) by computing every percentile from 1 to 99 inclusive. It should be clear that pupils who rank in the second quarter from the top of a distribution have scores between  $Q_3$  and the median and that those who rank in the middle fifth have scores between the fortieth and sixtieth percentiles. Similarly, pupils who have a percentile rank of, say, 37, have scores that lie between the thirty-seventh and thirty-eighth percentiles. It is apparent from the foregoing that a percentile is a *point* on the scale and that a percentile rank represents an *area* lying between two adjacent percentiles.

The teacher wishing to interpret test results is usually far more interested in percentile ranks than in percentiles, although there are occasions when he may wish to compute certain percentiles. The computation of percentiles will be treated briefly here, and a graphical method of value in the estimation of percentile ranks will be presented later in this chapter.

*Percentiles* are computed in the same manner as was illustrated in Table 15, page 325, for  $Q_1$  is the twenty-fifth percentile and  $Q_3$  is the seventy-fifth percentile. The median, for which computation procedures are shown in Table 15, is also a percentile—the fiftieth. The same 37 reading test scores frequently employed above are again used in the frequency distribution of Table 19. The cumulative



frequency column of the table is included as an aid in the computation of percentiles. Although percentiles near the top of a distribution are sometimes more easily computed by counting down from the top than by counting up from the bottom, the latter procedure will be employed in the illustrations here because of the manner in which the cumulative frequency column facilitates the computations.

TABLE 19. Computation of deciles and percentiles for the grouped frequency distribution of 37 reading test scores

<i>c.i.</i>	<i>f</i>	<i>cum. f</i>	
71-73	1	37	1. $\frac{N}{10} = \frac{37}{10} = 3.7$
68-70	2	36	2 cases in and below <i>c.i.</i> 29-31
65-67	1	34	$3.7 - 2 = 1.7$
62-64	1	33	$1.7 \div 2 = .85$
59-61	2	32	$.85 \times 3 = 2.55$
56-58	2	30	$31.50 + 2.55 = 34.05$ ( $P_{10}$ or $D_1$ )
53-55	3	28	2. $\frac{9}{10} N = .9 \times 37 = 33.3$
50-52	4	25	33 cases in and below <i>c.i.</i> 62-64
47-49	6	21	$33.3 - 33 = .3$
44-46	5	15	$.3 \div 1 = .33$
41-43	3	10	$.33 \times 3 = .99$
38-40	2	7	$64.50 + .99 = 65.49$ ( $P_{90}$ or $D_9$ )
35-37	1	5	3. $\frac{63}{100} N = .63 \times 37 = 23.31$
32-34	2	4	21 cases in and below <i>c.i.</i> 47-49
29-31	1	2	$23.31 - 21 = 2.31$
26-28	0	1	$2.31 \div 4 = .58$
23-25	1	1	$.58 \times 3 = 1.74$
<i>N</i>	37		$49.50 + 1.74 = 51.24$ ( $P_{63}$ )

Computations of three percentiles are illustrated in the table, but they will not be developed in detail here because they present nothing new to the student in the way of computational difficulties. Shown first in the table is the computation of the tenth percentile ( $P_{10}$ ), which is also known as the first decile ( $D_1$ ). The ninetieth percentile ( $P_{90}$ ), or ninth decile ( $D_9$ ), is shown in second position. These represent the procedures in computing deciles. The last illustration, to show procedures for percentiles in general, is for the sixty-third percentile ( $P_{63}$ ).

*Derived scores based on the standard deviation.* A considerable number of derived scores have the standard deviation and arithmetic

mean of a standard group of pupils as basic to their derivation. These various derived scores have different names, and some of them are devised for use with particular tests or series of tests. Although they differ widely in the manner in which the standard groups upon which they are based are selected, and make use of different numerical methods of representation, they have the element in common of being based upon the standard deviation.

Methods of computing the arithmetic mean and the standard deviation are presented in Chapter 12. One of their major uses is that of providing one of the most satisfactory means of deriving meaningful scores from test results. The brief treatment of derived scores here shows the major types of such scores and the elements of similarity and difference among them.

*Standard measures or z-scores* are mentioned briefly here because they represent such a simple method of showing deviation of a score from the arithmetic mean of the distribution and because of their similarity to other derived scores. However, the *z*-score is a measure used primarily in statistical procedures and has very little direct significance for the interpretation of test results to the teacher. The *z*-score is found by the application of the formula

$$z = \frac{X - M}{S.D.},$$

in which *X* is a particular raw score, *M* is the arithmetic mean of the distribution of raw scores, and *S.D.* is the standard deviation of the distribution of raw scores. It is sufficient here to point out that the *z*-score expresses deviation from the arithmetic mean in terms of standard deviation units and to give a few illustrations. For example, a *z*-score of +2.00 is two sigmas above the mean, a *z*-score of -2.00 is two sigmas below the mean, and a *z*-score of -.37 is .37 *S.D.* below the mean. Therefore deviations from the mean can be read directly from *z*-scores.

*T-scores* are similar to *z*-scores, except that they eliminate the use of negative values and decimals. A *T*-score of 50 was arbitrarily decided upon to represent a score at the arithmetic mean of a distribution and 10 *T*-score units were made equivalent to one standard deviation of distance. The formula for the *T*-score is

$$T = \frac{10(X - M)}{S.D.} + 50,$$



where  $X$ ,  $M$ , and  $S.D.$  have exactly the same significance as they had in computing  $z$ -scores, that is, a particular raw score, the arithmetic mean, and the standard deviation. A score two sigmas above the mean has a  $T$ -score value of 70, a score two sigmas below the mean has a  $T$ -score value of 30, and a score .37  $S.D.$  below the mean has a  $T$ -score equivalent of 46. Fractional values are not ordinarily used in  $T$ -scores.

*Standard scores, scaled scores, equated scores, and converted scores* are other types of derived scores that provide for comparability of scores on different parts of the same test or even on different tests. This is accomplished by changing raw scores to derived scores by methods differing somewhat from those described above but nevertheless based on the mean and standard deviation for some standard group.

## Other types of derived scores

Although the types of derived scores discussed here are those most commonly used, several miscellaneous types that do not fit into any of the categories above merit brief mention here.

In the field of intelligence testing, the personal constant ( $PC$ ) and the index of brightness ( $IB$ ) are not mentioned above, but they are given adequate treatment in Chapter 10. Personality inventories in a few instances make use of the personality quotient ( $PQ$ ), which is treated sufficiently in Chapter 11. A derived score that relates intelligence and achievement—the index of studiousness—is given attention in Chapter 10.

The derived scores discussed in this chapter and elsewhere in the volume probably do not include all of the types or variations of such measures, for it is not uncommon to find a new test appearing with a new type of derived score. However, the types presented are the most widely used and the most important at the present time.

## 3 INFORMAL TYPES OF DERIVED SCORES

Derived scores of the types discussed above are used variously in interpreting results from standardized tests, and particularly for standardized educational and intelligence tests. Some of these same derived scores can be, and in fact often are, used in interpreting results from teacher-made or classroom tests. This applies particu-

larly to percentile ranks but may also apply to *T*-scores. Three other methods of establishing comparability of results from teacher-made tests are discussed below: (1) relative ranks, (2) letter marks on a single test or measure, and (3) letter marks representing composite achievement covering a marking period, semester, or school year.

### Relative ranks

In working with test scores it often becomes desirable to make the achievement of all pupils in the group the basis for comparison. The relative performances of pupils may be compared by the simple process of assigning ranks or positions to their scores in accordance with the magnitudes of the scores. Thus for a group of twelve pupils who took a certain arithmetic test and received twelve different scores, ranks of 1, 2, 3, . . . 10, 11, and 12 are assigned in descending order of test scores.

The only difficulty appears when two or more pupils have identical scores. The illustration of Table 20 will make clear the method of handling this situation. Pupils B and C, with scores of 44, are assigned the average of the two rank positions—2 and 3—for which they are tied ( $2+3\div2$ ). Likewise, pupils F, G, and H, all with scores of 38, receive the rank of 7 ( $6+7+8\div3$ ).

TABLE 20. Assignment of relative ranks to arithmetic test scores

Pupil . . .	A	B	C	D	E	F	G	H	I	J	K	L
Score . . .	46	44	44	41	40	38	38	38	37	31	29	28
Rank . . .	1	2.5	2.5	4	5	7	7	7	9	10	11	12

It should be clear that the assignment of relative ranks or positions to pupils having certain test scores actually covers up the true situation to some degree. For example, the difference in magnitude of scores is submerged by relative ranks. To illustrate, in the data of Table 20 a difference of six score points on the test (31 to 37) makes a difference of only one rank position (9 to 10) in one situation. But a difference of only one score point (28 to 29) also makes a difference of one rank position elsewhere. This is a point that should be kept in mind when using the method of relative ranks. Ranking shows that one pupil is above or below another pupil but fails to indicate



how much in terms of actual score differences he is above or below the other pupil.

The usefulness of this method is also somewhat limited by the fact that it takes no account of the actual level at which the accomplishment takes place. A person ranking 32 in a group of 35 has a very low relative rank in the group. However, if he ranked 32 in a group of 250, the significance of his accomplishment would be greatly changed. Percentile ranks, however, take this point into account by reducing the ranking to a basis of 100 units. A percentile rank of 75 means that for the measures under consideration the individual made a higher score than 75 per cent of the individuals in his group without regard to the number of cases it contains.

### Letter marks on a test

One of the major uses of the standard deviation, as a measure basic to certain important types of derived scores, is treated in a preceding section of this chapter. The standard deviation has many other uses not directly related to formal derived scores, however, and one of them is important enough to justify attention here.

The student or teacher who is interested in the critical analysis of test scores will find the standard deviation a very useful measure in assigning letter marks to classroom test scores. The importance of this practice is sufficiently great that the steps involved in the technique are given in some detail. The computations described are based upon the 37 reading test scores used in the preceding computational illustrations and originally listed in Table 9, page 310. The steps of procedure used in assigning A, B, C, D, and F marks to the 37 pupils are outlined below by the use of the arithmetic mean and standard deviation of the scores.

1. *Obtain the arithmetic mean and standard deviation.* The arithmetic mean, shown in Table 14, page 320, is 48.96, and the standard deviation, shown in Table 18, page 334, is 10.86.
2. *Mark off distances of .5 S.D. above and below the mean.* The first of the points is  $48.96 + (10.86 \div 2)$ , or 54.39. The other point is  $48.96 - (10.86 \div 2)$ , or 43.53. These two points are respectively at the upper and lower ends of the C mark range.
3. *Mark off distances of 1.5 S.D. above and below the mean.* The first of these points is  $48.96 + (10.86 \times 1.5)$ , or 65.25. The other is

48.96 —  $(10.86 \times 1.5)$ , or 32.67. These two points separate the A and B and the D and F marks respectively.

4. *Establish the score limits of the letter marks.* From these values set up a table showing the score limits of each mark. It should be noted that no upper limit for the A mark and no lower limit for the F mark are set.

<i>Mark</i>	<i>Score Limits</i>
A	65.25 and above
B	54.39 to 65.25
C	43.53 to 54.39
D	32.67 to 43.53
F	32.67 and below

Reference to Table 10, page 310, where the 37 original scores are listed in descending order, will disclose that 3 A, 8 B, 16 C, 7 D, and 3 F marks are assigned by this method. The percentages of the 37 marks at each level are: A, 8; B, 22; C, 43; D, 19; and F, 8. These are close to the 38 per cent of Cs, 24 per cent each of Bs and Ds, and 7 per cent each of As and Fs which have a tendency to result if the distribution contains a rather large number of cases and if it approximates a normal distribution in form.

It is readily apparent that practically no subjective factors are involved in the assignment of marks by this method. The score limits are determined by the standard deviation units and would be the same no matter who assigned the marks. It should be noted, however, that these limits hold only for this particular distribution and must not be assumed to be true for any other test. The teacher should also remember that this method of marking does not take into account the absolute level of ability at which a particular class works. The superior pupil in an average or poor class receives an A by this method just as readily as does the superior pupil in a very superior class. This is probably less serious than it sounds, however, for most class groups large enough to warrant the application of this technique average out quite well in this respect.

### Final letter marks

Final marks summarizing all of the quiz and test scores, and even including all of the more subjective marks such as those on themes, term papers, and notebooks, can be obtained quite readily for the



work of a marking period, semester, or school year. Various methods for accomplishing this objectively have been employed. One of the best and simplest procedures involves the use of A, B, C, D, and F marks for stating results from each factor that is to receive consideration in determining the final marks. For valid and quite reliable measures, such as scores on carefully constructed objective tests, plus and minus marks in each letter category can be applied by a simple extension of the method outlined immediately above. For example, high C marks would be rated C+, low C marks would be assigned C-, and the intervening marks would be designated as C. For less reliable ratings, such as marks on themes and term papers, the five-point scale consisting of A, B, C, D, and F is doubtless preferable.

It is possible not only to use more discrimination in weighting highly reliable and valid scores than is used for more subjective measures, as is illustrated above, but also to weight the various factors entering into the determination of final marks according to their estimated importance. This is accomplished by using a weighting of 1 for least significant results, a weighting of 2 for results of intermediate importance, and a weighting of 3 for measures judged to be most important. Higher weights can easily be obtained if desired by an extension of Table 21.

TABLE 21. Suggested weightings for marks in obtaining composite scores

Mark	Weighting of		
	1	2	3
A+	16	32	48
A	15	30	45
A-	14	28	42
B+	12	24	36
B	11	22	33
B-	10	20	30
C+	8	16	24
C	7	14	21
C-	6	12	18
D+	4	8	12
D	3	6	9
D-	2	4	6
F	0	0	0

A simple illustration employing only three measures will suffice for showing how this method is applied. If a certain pupil has a B— on a mid-semester test, a C on his term paper, and a final examination mark of C+, and if the three measures are to be weighted 2, 1, and 3 respectively, his mid-semester test weighting is 20, his term paper weighting is 7, and his final examination weighting is 24. The sum of 20, 7, and 24, or 51, is his weighted composite score for total performance. When similar composites are obtained for the other pupils in the class, a distribution of the weighted composite scores can be made. It is then possible to assign final course marks by use of the method outlined above or by some modification of it. However, since no marking system should be rigidly defined, departures from any system of attaining objectivity should be made when conditions warrant.

### Exercises in Computing Derived Scores

13. Assign *T*-scores to the raw scores of 55, 66, 44, 79, and 35 from the 40 arithmetic test scores of Problem 1, using the values of the arithmetic mean and standard deviation found in Problem 4, page 323, and Problem 11, page 337, respectively.
14. Compute the 40th and the 63rd percentiles for the 40 arithmetic test scores tabulated in a frequency distribution for Problem 1, page 317.
15. Assign relative ranks to the 30 spelling test scores of Problem 3, page 317.
16. Determine limits for A, B, C, D, and F marks for the 40 arithmetic test scores of Problem 1, using the values of the arithmetic mean and standard deviation found in Problem 4, page 323, and Problem 11, page 337, respectively.
17. Find the weighted composite score for a pupil who has C+, B, C, and C+ marks respectively on a mid-semester test, a semi-final test, a term paper, and a final examination when the first two tests are weighted one each, the term paper is weighted two, and the final examination is weighted three.

## 4 GRAPHICAL REPRESENTATION

Various methods of graphical representation have value in the interpretation of results from educational measurements. These range from simple graphs and charts to complex and involved representations and to such popular techniques as the pictograph. Three types



of graphic representations seem to be most widely useful to the classroom teacher. Accordingly, only these three types will be treated here: (1) the frequency polygon, (2) the histogram, and (3) the cumulative frequency graph.

### Frequency polygon

The most widely used and the most easily constructed and comprehended form of graphical representation is the frequency polygon. When a normally distributed trait has been measured for a large number of pupils and when the data are grouped into a rather large number of class intervals, say 25 to 30, the resulting frequency

TABLE 22. Class frequencies, cumulative frequencies, and cumulative relative frequencies for 37 reading test scores

<i>c.i.</i>	<i>f</i>	<i>cum. f</i>	<i>cum. rel. f</i>
71-73	1	37	100.0
68-70	2	36	97.3
65-67	1	34	91.9
62-64	1	33	89.2
59-61	2	32	86.4
56-58	2	30	81.1
53-55	3	28	75.7
50-52	4	25	67.6
47-49	6	21	56.8
44-46	5	15	40.5
41-43	3	10	27.0
38-40	2	7	18.9
35-37	1	5	13.5
32-34	2	4	10.8
29-31	1	2	5.4
26-28	0	1	2.7
23-25	1	1	2.7
<i>N</i>	37		

polygon closely resembles in shape the smoothed normal curve pictured in Figure 28, page 331. A frequency polygon has major values in showing individual differences among the pupils in a certain class. Impressions concerning the range of scores, the shape of the distribution, and an approximation to a measure of the central tendency of the scores can readily be formed from this type of graph.

When certain points in the distribution, such as the median and quartiles, are designated with reference to the score scale, this type of graphical representation assumes much more meaning than does a frequency distribution of the same scores.

Reproduced in Table 22 is the grouped frequency distribution of the 37 reading test scores presented in their original form in Table 9, page 310, and presented in the form of a grouped frequency distribution in Table 13, page 314. Only the class-interval and frequency columns of this table are used in the construction of the frequency polygon; the right-hand columns will be used later in constructing a cumulative frequency graph for the same data. Figure 29 depicts this score distribution in the form of a frequency polygon. Values of the  $P_{10}$ ,  $Q_1$ , median,  $Q_3$ , and  $P_{90}$  are shown on the graph as an indication of how the polygon may be given additional meaning.

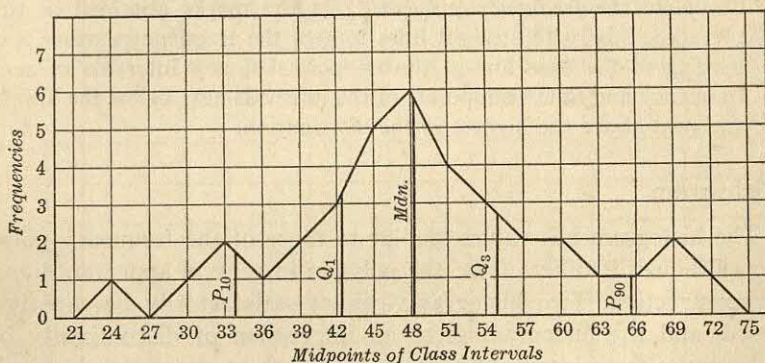


Fig. 29. Frequency polygon of 37 reading test scores

The steps of procedure for constructing a frequency polygon on squared paper as are follows:

1. *Rule left and bottom marginal lines.* Rule two straight lines, perpendicular to each other, to establish the left and bottom edges of the graph proper.
2. *Establish and indicate values on the score scale.* Lay off at equal distances along the base line and on rulings of the squared paper the midpoints of class intervals, starting somewhat to the right of the vertical line with the midpoint of the interval next below the lowest in the distribution and continuing toward the right to the midpoint of



the interval next above the highest in the distribution. Designate the values of these midpoints immediately below the base line.

3. *Establish and indicate values on the frequency scale.* Lay off on the left vertical line successive units to represent the frequencies of the different class intervals. Select a unit that will result in a height for the interval having the greatest frequency of perhaps two-thirds to three-fourths of the width of the figure. Designate values on the frequency scale immediately to the left of the vertical line.
4. *Complete the margins of the graph.* Rule a horizontal line across the top of the graph somewhat above the highest value on the frequency scale and rule a vertical line from a point somewhat to the right of the highest midpoint to complete the graph.
5. *Establish frequencies on the vertical scale.* At the midpoint of each interval along the base line, count up a distance along the frequency scale equal to the number of scores in the class interval and place a mark above the midpoint of the interval to indicate this height.
6. *Complete the frequency polygon.* Join the marks obtained in the preceding step with straight lines to give the frequency surface, extending to the base line at the midpoints of any intervals of zero frequency and at the midpoints of the intervals next below the lowest and next above the highest in the distribution.

## Histogram

The histogram has values similar to those of the frequency polygon, although it differs from the polygon in general appearance and in construction. Two histograms cannot satisfactorily be superimposed, and the histogram gives an impression of discontinuity of frequencies from interval to interval. It is the type of simple graph that closely resembles many of the pictographs in which silhouettes of figures are arranged in rows or columns of varying lengths to represent different frequencies. The frequency distribution of 37 reading test scores, reproduced in Table 22, is used to illustrate the construction of the histogram shown in Figure 30. Again the two right-hand columns of the table are not used in the procedure.

Although the construction of a histogram is similar to that of a frequency polygon, several minor differences in the early stages and major differences in the late stages warrant a separate list for the steps of procedure.

1. *Rule left and bottom marginal lines.* Follow the procedure given above in the first step for constructing a frequency polygon.

2. *Establish and indicate values on the score scale.* Lay off at equal distances along the base line and midway between rulings of the squared paper the midpoints of class intervals from left to right, starting somewhat to the right of the vertical line with the midpoint of the lowest interval and continuing to the midpoint of the highest interval. Designate the values of these midpoints immediately below the base line.
3. *Establish and indicate values on the frequency scale.* Follow the procedure given above in the third step for constructing a frequency polygon.
4. *Complete the margins of the graph.* Follow the procedure given above in the fourth step for constructing a frequency polygon.
5. *Establish frequencies on the vertical scale.* Follow the procedure given above in the fifth step for constructing a frequency polygon.
6. *Complete the histogram.* Rule horizontal lines from the lower to the upper limits of each class interval at the points where the marks of the last step of procedure were made. Complete the enclosure of the figure by ruling in the connecting vertical lines. Extend to the base line only at the upper and lower ends of the distribution and above and below class intervals having zero frequencies.

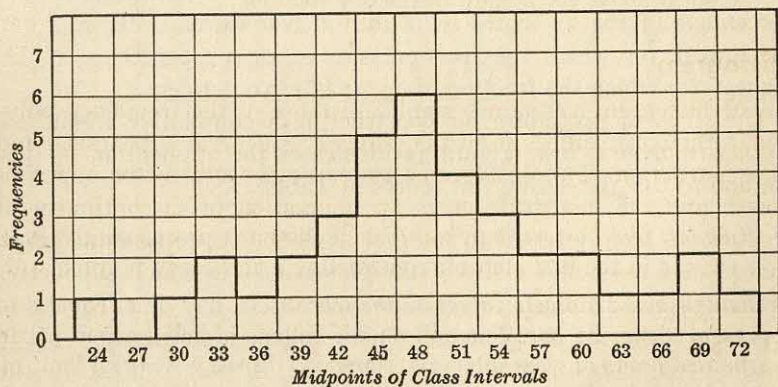


Fig. 30. Histogram of 37 reading test scores

### Cumulative frequency graph

This form of graphical representation does not reveal the major characteristics of a score distribution as clearly as do the frequency polygon and histogram. Therefore, it is not suitable for use in presenting test results graphically to point up individual differences



and group characteristics. This characteristic of the cumulative frequency graph results from the fact implied in its name—it shows cumulative frequencies rather than frequencies by class intervals. In doing so, it fails to reveal the shape of the curve and minor irregularities in it as well as do the frequency polygon and histogram.

The cumulative frequency graph has major significance in facilitating the estimation of quartiles, quintiles, deciles, and percentiles. This characteristic also makes possible the estimation of percentile ranks for given test scores. As has been shown above, percentile ranks are obtained by dividing the distribution into one hundred equal parts with respect to the area under the curve, i.e., in terms of score frequencies, by use of percentiles 1 to 100 inclusive. The 37 reading test scores used in the illustration here cannot readily be divided into that number of equal parts. Hence, cumulative relative frequencies are used in constructing the figure so that the frequency scale will be divided into the one hundred equal parts necessary for ready estimations of percentiles and percentile ranks. The last two columns of Table 22 show the 37 reading test scores cumulated upward by class intervals and the cumulated relative frequencies of the scores. The cumulative relative frequency column shows the percentage of the 37 scores lying in and below each interval from the lowest, for which the fraction is  $1/37$ , or 2.7 per cent, to the highest, for which the fraction of  $37/37$  is 100.0 per cent.

The steps of procedure for constructing a cumulative frequency graph are given below. Figure 31 illustrates the application of this method to the 37 reading test scores of Table 22.

1. *Rule left and bottom marginal lines.* Follow the procedure given on page 357 in the first step for constructing a frequency polygon.
2. *Establish and indicate values on the score scale.* Lay off at equal distances along the base line and on the rulings of the squared paper the real limits of class intervals, starting with the lower real limit of the lowest interval and continuing to the right until the higher real limit of the highest interval is reached. Designate the values of these real limits immediately below the base line.
3. *Establish and indicate values on the relative frequency scale.* Lay off on the left vertical line a scale in 100 parts to represent the cumulative relative frequencies. Select a unit such that the height of the graph will be roughly equal to or even somewhat greater than its width. Designate values on this vertical scale in multiples of ten or even of five.

4. *Complete the margins of the graph.* Rule a horizontal line starting at the 100 point on the vertical scale and a vertical line starting with the higher real limit of the highest interval in the distribution to complete the margin of the graph.
5. *Establish cumulative relative frequencies on the vertical scale.* Place a mark at the upper real limit of each class interval opposite the percentage on the vertical scale that shows the cumulative relative frequency in and below the interval.
6. *Complete the cumulative frequency graph.* Join these points successively by straight lines, starting with the point where the left vertical and base lines meet and continuing to the right and upward until the point where the right vertical and top horizontal lines meet is reached.

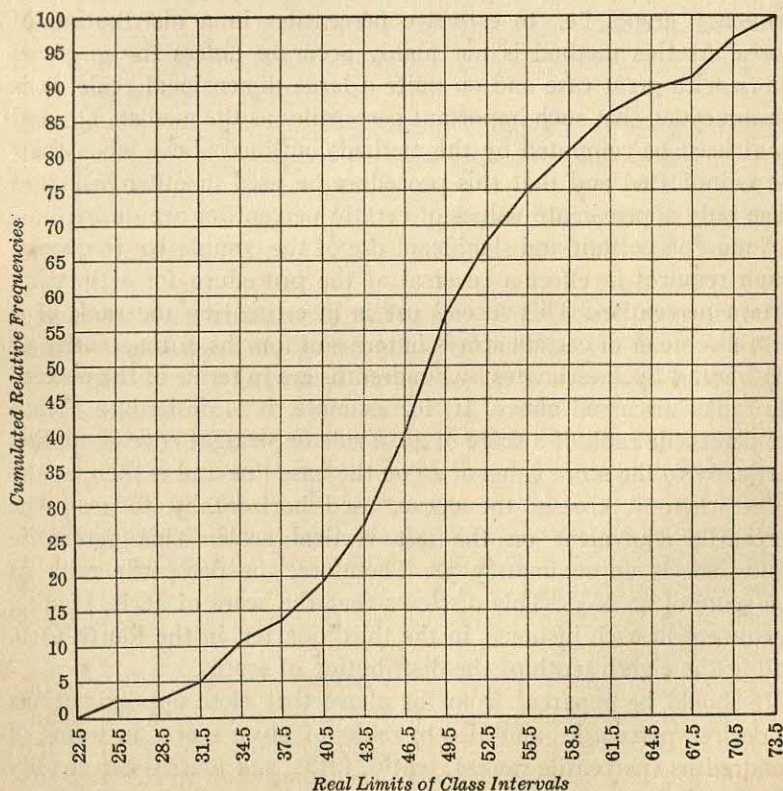


Fig. 31. Cumulative frequency graph of 37 reading test scores



It is suggested that the student use the cumulative frequency graph of Figure 31 to check the consistency of estimates based on it with results obtained above by computation for certain quartiles, deciles, and percentiles. A median of 48.24 was obtained for these 37 reading test scores in Table 15, page 325. Values of 65.49, 34.05, and 51.24 were given in Table 19, page 348, for  $D_{90}$ ,  $D_{10}$ , and  $P_{63}$  respectively. If, for example, it is desired to check on the twenty-fifth percentile, or  $Q_1$ , the procedure is to use a rule or straight edge horizontally in proceeding from the reading of 25 on the vertical scale to the curve itself and from that point on the curve to rule vertically downward with the straight edge to the score scale. The value on the score scale at this point should very closely approximate the 42.75 obtained computationally for  $Q_1$ .

The illustration above shows one of the uses of the cumulative frequency graph, i.e., to estimate percentiles in a distribution of scores. As this method is not highly accurate unless the graph is drawn with great care and on quite a large dimensional scale, it is recommended that such important percentiles as the median,  $Q_3$ , and  $Q_1$  at least be computed by the methods outlined above when their use is indicated and that this procedure be used in other instances when only approximate values of certain percentiles are desired.

A more important and significant use of the cumulative frequency graph requires in effect a reversal of the procedure for estimating certain percentiles. This second use is in estimating the rank of a certain score or of certain scores in terms of fourths, fifths, tenths, or hundredths. Such estimates by hundredths are in terms of the percentile ranks discussed above. If, for example, it is desired to obtain the percentile rank of a score of 46, a rule or straight edge is applied vertically to the score value of 46 on the base line and is then, at the point where it crosses the curve, used horizontally to read the percentile equivalent on the left vertical scale. This percentile equivalent is approximately 38. Therefore, the percentile rank of the score of 46 is 38. This discloses that the score of 46 is, reading downward in each instance, in the third quarter, in the fourth fifth, and in the eighth tenth of the distribution of scores.

It should be apparent from the above that close approximations to desired percentiles and also to ranks of given scores in terms of hundredths (percentile ranks), tenths, fifths, and fourths can readily be obtained from a carefully constructed cumulative frequency graph. The degree of accuracy in the estimates obtained is sufficient

for most practical purposes and the time and labor saved when a number of such values are to be computed is great.

### Exercises in Constructing Graphs

18. Construct a frequency polygon on squared paper for the 40 arithmetic test scores tabulated in a frequency distribution for Problem 1, page 317.
19. Construct a histogram on squared paper for the 30 language test scores tabulated in a frequency distribution for Problem 2, page 317.
20. Construct a cumulative frequency graph on squared paper for the 40 arithmetic test scores tabulated in a frequency distribution for Problem 1, page 317.

### Exercises in Estimating Percentiles and Percentile Ranks

21. Estimate the 40th and the 63rd percentiles for the 40 arithmetic test scores of Problem 1 from the cumulative frequency graph constructed for Problem 20.
22. Estimate the percentile ranks of raw scores of 37 and 56 in the 40 arithmetic test scores of Problem 1 from the cumulative frequency graph constructed for Problem 20.

## 5 MAJOR TYPES OF NORMS

In the interpretation of results from standardized tests, tables of norms nearly always have direct value. Norms of one of the three types discussed in Chapter 5 are ordinarily provided with such tests: (1) grade norms, (2) age norms, or (3) percentile norms. As has been pointed out in a previous section of this chapter, derived scores and norms overlap in various ways. In the section of this chapter devoted to derived scores, the various types of meaningful scores that can be derived either from tables of norms for standardized tests or from statistical manipulations of scores from standardized and informal objective tests were treated. Only those derived scores enter into the brief discussion of this section that are related to norms in one of two ways: (1) as results from the use of norm tables—grade scores, age scores, and percentile ranks, or (2) as scores intermediate between raw scores and final derived scores, e.g., standard scores, equated scores.

The tremendous variety of forms in which norm tables are presented for different standardized tests, whether of intelligence,



achievement, or personality, makes it impossible to represent here all of the variations found in such tables. Furthermore, the purpose here is only to familiarize the student sufficiently with the nature, form, and use of norms that he will be able to employ them properly in the interpretation of results from any standardized test he may have occasion to use. The discussion of norms for achievement tests in Chapter 5 and careful reading of instructions in test manuals should equip the teacher for effective use of such test norms as he is likely to encounter either in elementary or in high-school testing.

TABLE 23. Mental age norms for the Pintner General Ability Tests, Verbal Series <sup>1</sup>

Mdn. Stand. Score	Mental Age	Mdn. Stand. Score	Mental Age	Mdn. Stand. Score	Mental Age	Mdn. Stand. Score	Mental Age
100	6-11	125	9-1	150	12-0	175	15-7
101	7-0	126	9-2	151	12-1	176	15-9
102	7-1	127	9-3	152	12-2	177	16-0
103	7-2	128	9-5	153	12-3	178	16-2
104	7-3	129	9-6	154	12-5	179	16-4
105	7-4	130	9-7	155	12-6	180	16-6
106	7-5	131	9-9	156	12-8	181	16-9
107	7-6	132	9-10	157	12-9	182	17-0
108	7-7	133	9-11	158	12-11	183	17-2
109	7-8	134	10-0	159	13-0	184	17-4
110	7-9	135	10-1	160	13-2	185	17-7
111	7-10	136	10-2	161	13-3	186	17-10
112	7-11	137	10-4	162	13-5	187	18-0
113	8-0	138	10-5	163	13-7	188	18-3
114	8-1	139	10-7	164	13-9	189	18-6
115	8-2	140	10-8	165	13-10	190	18-8
116	8-3	141	10-9	166	14-0	191	18-11
117	8-4	142	10-10	167	14-3	192	19-2
118	8-5	143	11-0	168	14-4	193	19-5
119	8-6	144	11-2	169	14-6	194	19-8
120	8-7	145	11-3	170	14-8	195	19-11
121	8-8	146	11-5	171	14-10	196	20-2
122	8-9	147	11-7	172	15-0	197	20-5
123	8-10	148	11-8	173	15-2	198	20-8
124	9-0	149	11-10	174	15-5	199	21-0

<sup>1</sup> *Directions for Administering and Scoring: Pintner General Ability Tests, Intermediate and Advanced.* World Book Co., Yonkers, N. Y., 1938. Table 1, p. 5.

TABLE 24. Grade and age norms for the California Language Test<sup>2</sup>

Grade Place- ment	Mech. Eng. and Gram.	Spell- ing	Total Lan- guage	Age in Months	Grade Place- ment	Mech. Eng. and Gram.	Spell- ing	Total Lan- guage	Age in Months
3.5	10	1	10-11	105	7.5	41	—	54	154
3.6	11	—	12	106	7.6	—	14	55	155
3.7	12	—	13	107-8	7.7	42	—	56	156-7
3.8	13	—	14-15	109	7.8	—	—	57	158
3.9	14	—	16	110	7.9	43	15	58	159
4.0	15	2	17	111	8.0	44	—	59	160
4.1	16	—	18-19	112	8.1	—	16	60	161
4.2	17	—	20	113-4	8.2	45	—	61	162-3
4.3	18	—	21	115	8.3	—	—	62	164
4.4	19	—	22	116	8.4	46	17	63	165
4.5	20	3	23	117	8.5	47	—	64	166
4.6	21	—	24-25	118	8.6	—	18	65	167
4.7	22	—	26	119	8.7	48	—	66	168-9
4.8	—	4	27	120-1	8.8	—	—	67	170
4.9	23	—	28	122	8.9	49	19	68	171
5.0	24	5	29	123	9.0	—	—	—	172
5.1	—	—	30	124	9.1	50	—	69	173
5.2	25	—	31	125-6	9.2	—	20	70	174-5
5.3	—	6	32	127	9.3	51	—	71	176
5.4	26	—	33	128	9.4	—	21	72	177
5.5	27	7	34	129	9.5	52	—	73	178
5.6	—	—	35	130-1	9.6	—	22	74	179
5.7	28	—	36	132	9.7	53	—	75	180-1
5.8	29	8	37	133-4	9.8	—	—	76	182
5.9	30	—	38	135	9.9	54	23	77	183
6.0	31	—	39	136	10.0	55	—	78	184
6.1	—	9	40	137	10.1	—	24	79	185
6.2	32	—	41	138-9	10.2	—	—	80	186-7
6.3	—	—	42	140	10.3	56	25	81	188
6.4	33	—	43	141	10.4	—	—	—	189
6.5	34	10	44	142	10.5	57	—	82	190
6.6	—	—	45	143	10.6	—	—	83	191
6.7	35	—	46	144-5	10.7	58	26	84	192-3
6.8	36	11	47	146	10.8	—	—	85	194
6.9	37	—	48	147	10.9	59	—	86	195
7.0	—	12	49	148	11.0	60	27	87-88	196
7.1	38	—	50	149	11.5	61	28	89	201
7.2	39	—	51	150-1	12.0	62	—	90	207
7.3	—	13	52	152	12.5	63	29	91-92	213
7.4	40	—	53	153	13.0	64	—	93	219

<sup>2</sup> Ernest W. Tiegs and Willis W. Clark, *Manual for the California Language Test*, Intermediate. California Test Bureau, Los Angeles, 1950. p. 20.



TABLE 25. Percentile norms for the Cooperative Mechanics of Expression Test, A<sup>3</sup>

End-of-Year Norms in Terms of Scaled Scores						
Scaled Score	Grade					
	7	8	9	10	11	12
				Percentiles *		
76						99
74						98
72					99	97
70					98	95
68				99	97	93
66				98	95	90
64			99	97	93	86
62			98	95	89	82
60		99	97	92	85	76
58		98	95	89	79	69
56		97	93	84	73	61
54	99	95	89	78	65	53
52	98	93	84	71	56	44
50	96	89	78	62	48	36
48	94	84	70	54	39	29
46	90	77	62	45	31	22
44	86	70	53	36	24	17
42	79	61	44	28	18	13
40	72	52	35	21	13	9
38	63	43	27	15	9	6
36	54	34	20	10	6	3
34	44	26	14	7	4	2
32	35	19	10	4	2	1
30	27	13	6	3	1	
28	19	9	4	2		
26	13	6	2	1		
24	9	3	1			
22	6	2				
20	3	1				
Mean	35.2	39.6	43.4	47.2	50.5	53.3
S.D.	8.3	8.6	8.7	8.9	9.2	9.5

\* The percentile values in the tables are those closest to the actual Scaled Scores listed. Interpolation may be used to obtain the closest percentiles for odd-numbered Scaled Scores.

<sup>3</sup> *Secondary School Norms: Cooperative English Test, Single Booklet Edition, Higher and Lower Levels. Cooperative Test Division, Educational Testing Service, New York. Norms for Public Secondary Schools of the East, Middle West, and West, p. 2.*

TABLE 26. Percentile norms for the Aspects of Personality Inventory <sup>4</sup>  
For Grades 7, 8, and 9

ANY TEST SCORE	PERCENTILE RANK CORRESPONDING TO GIVEN SCORE						ANY TEST SCORE
	BOYS			GIRLS			
	Sec. I A-S	Sec. II E-I	Sec. III E	Sec. I A-S	Sec. II E-I	Sec. III E	
35						100	35
34			100			97	34
33			92			93	33
32			75		100	83	32
31		100	66		99	70	31
30		99	53		99	58	30
29		99	42		97	49	29
28	100	98	36	100	96	39	28
27	99	97	27	99	94	34	27
26	98	93	18	99	88	27	26
25	97	83	13	98	81	23	25
24	96	79	12	97	72	18	24
23	95	69	10	94	63	15	23
22	90	55	9	90	53	11	22
21	88	42	8	82	43	9	21
20	77	32	8	77	30	7	20
19	71	24	6	72	20	6	19
18	63	22	5	63	12	4	18
17	55	13	4	56	11	3	17
16	42	9	3	48	9	3	16
15	37	6	2	38	6	2	15
14	27	5	1	27	5	2	14
13	19	3	1	25	4	2	13
12	15	2		19	3	1	12
11	6	1		14	2	1	11
10	5	1		10	2	1	10
9	3			4	1		9
8	1			3	1		8
7	1			2			7
6				1			6

<sup>4</sup> Rudolf Pintner and others, *Aspects of Personality: Manual of Directions*. World Book Co., Yonkers, N. Y., 1939. Table 5, p. 7.



## Exercises in Using Test Norms

23. Using the mental age norms reproduced in Table 23 for the *Pintner General Ability Tests, Verbal Series*, determine the mental ages and, using the quotient method, the intelligence quotients of the pupils whose median standard scores and chronological ages are shown below.

Pupils	Median Standard Scores	Chronological Ages	Pupils	Median Standard Scores	Chronological Ages
A	172	15-0	D	165	15-0
B	105	6-8	E	182	12-9
C	107	8-6	F	133	10-2

24. Using the grade placement and age norms reproduced in Table 24 for the *California Language Test*, determine the grade equivalents and age equivalents (in years and months) on the two parts and the total test for the eighth-grade pupils whose scores are shown below.

Test	Pupils					
	A	B	C	D	E	F
Mechanics of English and Grammar	44	50	42	15	55	49
Spelling	17	18	16	3	21	13
Total Language	61	68	58	18	76	62

25. Using the percentile norms reproduced in Table 25 for the *Co-operative English Mechanics Test*, determine the percentile rank in his grade group for each of the pupils whose end-of-year scaled scores are shown below.

Pupils	Grades	Scaled Scores	Pupils	Grades	Scaled Scores
A	9	54	D	10	51
B	7	35	E	8	64
C	11	41	F	9	27

26. Using the percentile norms reproduced in Table 26 for the *Aspects of Personality* inventory, determine the percentile rank on each section of the test for the pupils whose scores and sexes are shown below.

Pupils	Sexes	Section Scores		
		Ascendancy-Submission	Introversion-Extroversion	Emotional Stability
A	Boy	25	19	31
B	Girl	15	25	30
C	Boy	18	22	29
D	Girl	22	18	26
E	Girl	17	21	29
F	Boy	14	24	27

### Selected References

- BROOM, M. E. *Educational Measurements in the Elementary School*. New York: McGraw-Hill Book Co., Inc., 1939. p. 44-48, 62-68.
- CONRAD, HERBERT S. "Comparable Measures." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 279-82.
- CONRAD, HERBERT S. "Norms." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 795-802.
- DARLEY, JOHN G. *Testing and Counseling in the High-School Guidance Program*. Chicago: Science Research Associates, 1943. p. 57-63.
- FLANAGAN, JOHN C. "Units, Scores, and Norms." *Educational Measurement*. Washington, D. C.: American Council on Education, 1951. Chapter 17.
- GARRETT, HENRY E. *Statistics in Psychology and Education*. Fourth edition. New York: Longmans, Green and Co., 1953. p. 9-20; Chapter 4.
- GERBERICH, J. RAYMOND. "Conversion of Scores." *Encyclopedia of Modern Education*. New York: Philosophical Library, 1943. p. 715-16.
- GREENE, EDWARD B. *Measurements of Human Behavior*. Revised edition. New York: Odyssey Press, 1952. p. 356-63, 372-77.
- GREENE, HARRY A., AND CRAWFORD, JOHN R. *Work-Book in Educational Measurements and Evaluation*. New York: Longmans, Green and Co., 1945. Units 7-9.
- LEE, J. MURRAY. *A Guide to Measurement in Secondary Schools*. New York: D. Appleton-Century Co., Inc., 1936. Chapter 7.



- LINCOLN, EDWARD A., AND WORKMAN, LINWOOD L. *Testing and the Use of Test Results*. New York: Macmillan Co., 1935. p. 94-97; Chapter 6.
- LINDQUIST, E. F. *A First Course in Statistics*. Revised edition. Boston: Houghton Mifflin Co., 1942. Chapters 3-4, 9.
- MICHEELS, WILLIAM J., AND KARNES, M. RAY. *Measuring Educational Achievement*. New York: McGraw-Hill Book Co., Inc., 1950. p. 414-23, 436-39, 442-51.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. Chapter 14.
- ODELL, C. W. *An Introduction to Educational Statistics*. New York: Prentice-Hall, Inc., 1946. Chapters 3, 5.
- REMMERS, H. H., AND GAGE, N. L. *Educational Measurement and Evaluation*. New York: Harper and Brothers, 1943. p. 526-37, 550-58.
- RICHARDSON, M. W. "The Logic of Age Scales." *Educational and Psychological Measurement*, 1:25-34; January 1941.
- ROSS, C. C. *Measurement in Today's Schools*. Second edition. New York: Prentice-Hall, Inc., 1947. Chapters 9-10, 14.
- SMITH, EUGENE R. "Recording for Guidance and Transfer." *Appraising and Recording Student Progress*. New York: Harper and Brothers, 1942. Part II.
- TABA, HILDA. "Interpretation and Uses of Evaluation Data." *Appraising and Recording Student Progress*. New York: Harper and Brothers, 1942. Chapter 7.
- TIEGS, ERNEST W. *Tests and Measurements in the Improvement of Learning*. Boston: Houghton Mifflin Co., 1939. Chapter 18.
- TRAXLER, ARTHUR E. *Techniques of Guidance: Tests, Records, and Counseling in a Guidance Program*. New York: Harper and Brothers, 1945. Chapter 13.
- TRAXLER, ARTHUR E. AND OTHERS. *Introduction to Testing and the Use of Test Results*. New York: Harper and Brothers, 1953. Chapter 7.
- WALKER, HELEN M. *Elementary Statistical Methods*. New York: Henry Holt and Co., 1943. p. 36-46; Chapter 5.
- WALKER, HELEN M. *Mathematics Essential for Elementary Statistics*. New York: Henry Holt and Co., 1934.
- WEITZMAN, ELLIS, AND MCNAMARA, WALTER J. *Constructing Classroom Examinations*. Chicago: Science Research Associates, 1949. p. 108-12; Chapter 12.

## ***Determining Relationships among the Results of Measurement***

THE FOLLOWING points involving the relationships among test scores are discussed in this chapter :

- A. Need for measures of relationship.
- B. Correlation coefficients as measures of relationship.
- C. Computation of the product-moment correlation coefficient.
- D. Meaning of the correlation coefficient.
- E. Determination of test validity.
- F. Determination of test reliability.

The discussion of Chapter 12 was concerned with the classification of test scores and computation of the two basic types of measures used in describing a single set of scores—measures of central tendency and measures of variability or dispersion. Chapter 13 presented various types of formal and informal derived scores, graphical representation, and norms, all designed to give meaning to a single set of scores. There remains to be considered the type of situation in which two sets of scores are obtained for the same group of individuals and in which some measure of the relationship between the two sets of scores is desired.

### **1 RELATIONSHIP BETWEEN SETS OF TEST SCORES**

#### **Need for measures of relationship**

In the selection, construction, and use of educational measuring instruments there are many situations in which a reasonably exact



expression of the relationship existing between two sets of measures is necessary. For example, the one test that most nearly measures the desired ability must be selected from a series of related tests. The method followed in such a case involves finding the relationship between the several tests and the ability to be measured. This procedure, called the method of *correlation*, is applied when two, or even more than two, measures of the same individuals are employed in determining the degree to which certain tested traits or abilities are related. In practical test construction this method is used in obtaining estimates of the validity, reliability, and objectivity of a test.

### Nature of the correlation coefficient

In the expression of relationships, as in other statistical measures, it is desirable to use a single mathematical value. Methods have been developed for describing relationships in terms of the correspondence between rank positions of scores and in terms of the percentage of scores falling within a specified unit of variability of each other, but ordinarily these procedures lack sufficient exactness to warrant their general use in the analysis of test results. The student who is interested in these different methods will find them discussed in certain of the treatments on statistical methods listed in the references at the end of this chapter. The one method considered here, the *Pearson Product-Moment Method*, is by far the most common and is, on the whole, the basic method used in educational investigations. This method, while somewhat complicated and difficult because of the large number of different calculations to be made, really involves comparatively little that is new to the student.

*The Pearson product-moment coefficient of correlation, indicated by  $r$ , is a single numerical index that expresses the extent to which the pairs of corresponding measures of two variables tend to deviate similarly from their respective arithmetic means.* The values of  $r$  may vary all the way from  $+1.00$ , indicating perfect positive relationship, through all of the possible decimals to zero ( $0.00$ ), indicating no relationship whatever, to  $-1.00$ , indicating a perfect negative relationship. The following are illustrations of a positive relationship between two factors:

- (1) The rise and fall of a column of mercury in a thermometer with the rise and fall of the outside temperature. As the temperature rises the column of mercury also rises.

- (2) The direction of the wind and the movement of smoke from a chimney. The smoke moves away with the wind.
- (3) The tendency of pupils who are intelligent to be good silent readers.

Negative correlation may be illustrated by:

- (1) The movement of the elevator cage and the counterbalancing weights. As the elevator cage goes up, the counterbalancing weights move in the opposite direction.
- (2) The relation between absence from school and school achievement.

Zero or indifferent correlation is best illustrated by means of the chance matching of numbered cards that have been shuffled. Two packs of 25 blank cards each may be numbered and the packs carefully shuffled so that the cards are in no systematic order. If cards are drawn at random from each pack and paired, the resulting relationship is likely to be close to zero. If these same packs of cards are both arranged in ascending order and the first card from one paired with the first card from the other, the resulting relationship will be positive. If one pack is inverted and each time a small-numbered card is taken from the one pack a large-numbered card is taken from the other, the resulting correlation will be negative.

This illustration of the numbered cards suggests one of the simple methods of expressing correlation, *viz.*, the method of ranking. If pupils are given two tests and the scores from the tests tend to place the same pupils in the same relative positions in each series, there is an indication of a positive correlation between the two tests. For example, the accompanying pairs of scores for nine pupils indicate a high positive relationship between the two tests because the pupil making the highest score on Test A also made the highest score on Test B and each other pupil in the list maintained his relative position on both tests. This suggests that the two tests measure abilities which have a great many factors in common. On the other hand, if it had happened that Pupil 1, who made a score of 89 on Test A had made a score of 18 on Test B, Pupil 2, who made a score of 85 had made a score of 20 on Test B, and scores for the other pupils were similarly interchanged, the resulting correlation would have been negative and would have shown that the two tests were inversely related. That is, the negative correlation would have indicated that high ability on Test A accompanied low ability on Test B.

The Pearson product-moment coefficient of correlation is computed from data arranged in a frequency table, but the table used is in a



different form from that employed in any of the problem work in this book thus far. The method of tabulating paired data is explained in the following section.

TABLE 27. Pairs of test scores

Pupil Number	Score on	
	Test A	Test B
1	89	32
2	85	31
3	83	29
4	80	28
5	76	26
6	70	24
7	65	21
8	61	20
9	54	18

## 2 COMPUTATION OF PEARSON PRODUCT-MOMENT CORRELATION COEFFICIENT

Although the Pearson product-moment coefficient of correlation is not always the easiest measure of relationship to compute, it is the most reliable and the most widely used measure of this type. A detailed illustration of its computation is given on the following pages.

### Computing the Pearson product-moment $r$

The speed and comprehension scores made by 90 pupils on a certain reading test are used in the following illustration showing how the scores are tabulated in a double-entry table and the correlation coefficient is computed. The procedures are illustrated in Table 28.

(1) *Set up a double-entry table.* Construct a frequency distribution for one set of scores along the left side of the table in form identical to that used in setting up a single frequency distribution in Chapter 12. This distribution is on the  $Y$ -axis of the correlation chart. Construct a similar frequency distribution for the second set of test scores horizontally across the top of the table with the low scores at the left and the high scores at the right. This distribution is on the  $X$ -axis of the chart. Extend the chart to the right by adding

columns headed  $f_y$ ,  $d_y$ ,  $fd_y$ ,  $fd_y^2$ , and  $xy$ . Similarly extend the chart downward by adding rows headed  $f_x$ ,  $d_x$ ,  $fd_x$ , and  $fd_x^2$ . It is desirable to use squared paper in setting up the correlation table. The result should have the same general form as does Table 28.

(2) *Tabulate the pairs of scores in the double-entry table.* A tally mark in a frequency distribution shows one fact—the position in the distribution of the score tabulated. A tally mark in a double-entry table or scatter diagram shows two facts—the score made by the pupil on each of two tests. Therefore, tabulate the scores so that the tally for each pair of scores is in the table cell that simultaneously represents the score of the individual on each of the two tests. For example, the tally mark in the lower left corner of Table 28 accounts for a speed score of 9 and a comprehension score of 21 made by a certain pupil. The mark is in the row for speed scores of 8 to 10 and in the column for comprehension scores from 18 to 22. For another pupil, having speed and comprehension scores of 32 and 49 respectively, a tally mark appears in the cell where the 32-34 row and the 48-52 column cross. The remaining 88 scores, not shown here, were similarly tabulated in the appropriate cells to complete the scatter diagram of the table.

In completing the scatter diagram, tabulate each pair of scores separately and total the tallies in each row and in each column, recording the resulting frequencies in the  $f_y$  column and the  $f_x$  row. Separately add the frequencies in the column and row to obtain the total number of cases. The two totals should agree exactly, as shown in the  $N$  of 90 for Table 28.

(3) *Assume values for the means and count off the deviations.* Assume a value for the mean of the scores on the  $Y$ -axis and count off the deviations in the  $d_y$  column, as was done in step 2, page 322, in the computation of the arithmetic mean. Similarly, assume a mean for the distribution of the  $X$ -axis and count off deviations in the  $d_x$  row to the right (positive signs) and to the left (negative signs) of the interval in which the mean is assumed to lie.

In Table 28, since the mean of the speed scores on the  $Y$ -axis was assumed to be 33.00, or the midpoint of the interval 32-34, the deviations were counted upward and downward from that interval. Likewise, since the mean of the comprehension scores on the  $X$ -axis was assumed to be 45.00, or the midpoint of the interval 43-47, the deviations were counted to the right (positive) and to the left (negative) from that interval.



TABLE 28. Computation of the Pearson product-moment coefficient of correlation between speed and comprehension scores on a certain reading test

Comprehension

	18	23	28	33	38	43	48	53	58	$f_y$	$d_y$	$fd_y$	$fd_y^2$	$xy$
	22	27	32	37	42	47	52	57	62					
53-55									2	2	+7	+14	98	42
50-52										0	+6	0	0	0
47-49									4	4	+5	+20	100	60
44-46								1	4	5	+4	+20	80	56
41-43							4	1	2	7	+3	+21	63	36
38-40							4	1	3	8	+2	+16	32	30
35-37						5	2	1		8	+1	+8	8	4
32-34					2	5	1		1	9	0	0	0	0
29-31				4	2	3				9	-1	-9	9	10
26-28			2	4	5	1	1			13	-2	-26	52	36
23-25			4	5	3					12	-3	-36	108	75
20-22		3	1	1	2	1				8	-4	-32	128	76
17-19		1		1		1				3	-5	-15	75	30
14-16			1							1	-6	-6	36	18
11-13										0	-7	0	0	0
8-10	1									1	-8	-8	64	40
$f_x$	1	4	8	15	14	16	12	4	16	90		-33	853	513
$d_x$	-5	-4	-3	-2	-1	0	+1	+2	+3					
$fd_x$	-5	-16	-24	-30	-14	0	+12	+8	+48	-21				
$fd_x^2$	25	64	72	60	14	0	12	16	144	407				

$$\frac{\Sigma fd_x}{N} = \frac{-89 + 68}{90} = \frac{-21}{90} = -.233 (c_x)$$

$$\frac{\Sigma fd_y}{N} = \frac{-132 + 99}{90} = \frac{-33}{90} = -.367 (c_y)$$

$$\Sigma fd_x^2 = \frac{407}{90} = 4.522$$

$$\Sigma fd_y^2 = \frac{853}{90} = 9.478$$

$$c_x^2 = -.233^2 = .054$$

$$c_y^2 = -.367^2 = .135$$

$$\sigma_x = \sqrt{4.522 - .054} = \sqrt{4.468} = 2.114$$

$$\sigma_y = \sqrt{9.478 - .135} = \sqrt{9.343} = 3.057$$

$$\frac{\Sigma xy}{N} = \frac{513}{90} = 5.700$$

$$c_x c_y = -.233 \times -.367 = .086$$

$$\sigma_x \sigma_y = 2.114 \times 3.057 = 6.462$$

$$r = \frac{\frac{\Sigma xy}{N} - c_x c_y}{\sigma_x \sigma_y} = \frac{5.700 - .086}{6.462} = \frac{5.614}{6.462} = +.869$$

(4) *Compute the standard deviations in class-interval form.* Compute the standard deviations for the  $Y$ -axis and the  $X$ -axis scores by the procedure outlined in Chapter 12, except that the final step of multiplying by the size of the class interval is omitted. Since the entire process of computing the correlation coefficient is carried on by the use of the class-interval scales rather than the score scales, the standard deviations are here stated in class-interval rather than in score units. Except for the fact that the  $f$ ,  $d$ ,  $fd$ , and  $fd^2$  values for the  $X$  distribution appear across the table at the bottom, rather than vertically as for the  $Y$  distribution, the steps in computing the standard deviations present no new difficulties. The two standard deviations,  $\sigma_x$  and  $\sigma_y$ , with the subscripts indicating the  $X$ -axis and  $Y$ -axis data, are 2.114 and 3.057, respectively, for the data of Table 28.<sup>1</sup>

(5) *Compute the sum of the product moments.* The name "product-moment method" implies the significant feature of the process. The relationship itself takes into account the operation of forces (frequencies) at varying distances (deviations in intervals) from the point of rotation (mean) on each axis. Since each measure assumes a position with regard to each of the two axes, the resulting moments must take this fact into account.<sup>2</sup>

Table 29 is presented both to illustrate the principle of product moments and as an aid to the student in later computations. As the 0 row on the  $Y$ -axis and the 0 column on the  $X$ -axis represent the assumed means of the two distributions, deviations are counted in both directions from the 0 row and the 0 column in the same manner as is shown in Table 28. The moment of any cell in the table is the product of its deviations on the two axes when the signs of the deviations are taken into account. For example, the moment of the cell in the upper right corner (21) is obtained as the product of its two deviations— $7 \times 3 = 21$ . Similarly, the moment of the cell having a deviation of +1 on the  $X$ -axis and -2 on the  $Y$ -axis is -2 ( $1 \times -2 = -2$ ). The moment of each cell is given in its upper right corner.

<sup>1</sup> Computations for the correlation coefficient should uniformly be carried to four decimal places and rounded back to three decimal places, otherwise using the same procedure for rounding numbers as was illustrated in Chapter 12, page 321.

<sup>2</sup> The reader will recall the illustration of moments of forces given in Chapter 12 in connection with the computation of the arithmetic mean. Forces are brought into balance by equating their moments, whether two forces are operating in one direction, as is the case for a single frequency distribution, or whether two forces are operating in each of two different directions, as is the case for a scatter diagram.



TABLE 29. Moments of cells in a double-entry table

 $X = \text{Axis}$  $Y = \text{Axis}$ 

-35	-28	-21	-14	-7		7	14	21	+ 7
-30	-24	-18	-12	-6		6	12	18	+ 6
-25	-20	-15	-10	-5		5	10	15	+ 5
-20	-16	-12	- 8	-4		4	8	12	+ 4
-15	-12	- 9	- 6	-3		3	6	9	+ 3
-10	- 8	- 6	- 4	-2		2	4	6	+ 2
- 5	- 4	- 3	- 2	-1		1	2	3	+ 1
									0
5	4	3	2	1		-1	- 2	- 3	- 1
10	8	6	4	2		-2	- 4	- 6	- 2
15	12	9	6	3		-3	- 6	- 9	- 3
20	16	12	8	4		-4	- 8	-12	- 4
25	20	15	10	5		-5	-10	-15	- 5
30	24	18	12	6		-6	-12	-18	- 6
35	28	21	14	7		-7	-14	-21	- 7
40	32	24	16	8		-8	-16	-24	- 8
- 5	- 4	- 3	- 2	- 1	0	+ 1	+ 2	+ 3	

Values carried to the  $xy$  column at the right side of Table 28 are sums of the product moments for all scores in each row of the table. Since the two scores in the upper right corner have moments of 21 each, the product moment of 42 ( $2 \times 7 \times 3$ ) is shown in the  $xy$  column. The 13 scores in the interval 26-28 on the speed test will be used for an additional illustration. The two scores at the left have a moment of 6 each ( $-2 \times -3$ ); thus their product moment is 12 ( $2 \times 6$ ). The product moment, shown in one operation, for the next four scores is 16 ( $4 \times -2 \times -2$ ). Similarly, the next five scores have a product moment of 10 ( $5 \times -2 \times -1$ ). The next score is in the 0 column on the  $X$ -axis, and consequently has a moment of 0 ( $1 \times -2 \times 0$ ). Therefore, the sum of the positive product moments, as obtained above, is 38 ( $12 + 16 + 10$ ). But the remaining score, to the right of the 0 column, must also be taken into account. Its product moment is  $-2$  ( $1 \times -2 \times 1$ ). The algebraic sum of the 38 and the  $-2$  is 36; hence, that is the  $xy$  value shown in Table 28 for the 13 scores in the  $Y$ -axis interval 26-28. The other  $xy$  values were similarly computed, and the sum of these values, or  $\Sigma xy$ , was found to be 513.

In computing the product moments, follow the procedure illustrated above. First find the moment of each cell in which at least one tally mark lies. Then find the product moment of each cell. Next sum the product moments, taking account of signs, in each row of the table and carry the results to the  $xy$  column. Finally, add the  $xy$  values to obtain  $\Sigma xy$ .

(6) *Obtain the mean of the product moments.* Divide the  $\Sigma xy$  of the preceding step by the number of cases to obtain  $\Sigma xy/N$ . For the data of Table 28, this is  $513/90$  or  $5.700$ .

(7) *Obtain the product of the corrections.* As  $\Sigma fd/N = c$ , both on the  $X$ -axis and the  $Y$ -axis, multiply  $c_x$  by  $c_y$ . For Table 28, these values are  $-.233$  and  $-.367$ , and their product was found to be  $.086$ .

(8) *Obtain the product of the standard deviations.* Obtain the product of the standard deviations on the two axes. Since the values given in Table 28 are  $2.114$  ( $\sigma_x$ ) and  $3.057$  ( $\sigma_y$ ), their product ( $\sigma_x \sigma_y$ ) was found to be  $6.462$ .

(9) *Substitute in the formula and solve for  $r$ .* Obtain the correlation coefficient by substituting the values obtained in the above steps in the formula:

$$r = \frac{\frac{\Sigma xy}{N} - c_x c_y}{\sigma_x \sigma_y},$$



in which  $r$  is the correlation coefficient,  $N$  is the number of cases,  $c_x$  and  $c_y$  are corrections on the  $X$ -axis and the  $Y$ -axis,  $\sigma_x$  and  $\sigma_y$  are the standard deviations on the  $X$ -axis and the  $Y$ -axis, and  $\Sigma xy$  is the sum of the product moments for all scores in the scatter diagram. For the correlation chart of Table 28, the numerator of the fraction is  $5.700 - .086$ , or  $5.614$ . The denominator, given above, is  $6.462$ . Therefore,  $r$  is  $5.614 \div 6.462$ , or  $+.869$ .

### Summary of steps in computing the Pearson product-moment correlation coefficient

Given below in summary form are the steps of procedure for computing the Pearson product-moment  $r$  by the use of a double-entry table or scatter diagram.

- (1) Set up a double-entry table. Construct a frequency distribution at the left of the chart ( $Y$ -axis) for one of the variables, using the procedure of steps 1 to 3, page 316, for setting up a frequency distribution. Construct a second frequency distribution by the same procedure across the top of the chart ( $X$ -axis) for the other variable, with the low scores at the left and the high scores at the right.
- (2) Tabulate the pairs of scores in the double-entry table. Place a tally mark in the cell of the table that correctly represents the paired scores of each individual on the  $Y$ -axis and  $X$ -axis variables. Sum the tally marks in the  $f_y$  column to obtain  $N$  and in the  $f_x$  row to obtain  $N$  again as a check on accuracy.
- (3) Assume values for the means and count off the deviations. Assume a value for the mean on each axis and count off the deviations in the  $d$  column and row, using the procedure of steps 1 and 2, page 322, for computing the arithmetic mean. Deviations upward and to the right are positive; deviations downward and to the left are negative.
- (4) Compute the standard deviations in class-interval form. Obtain for both the  $X$ -axis and the  $Y$ -axis data the necessary sums of the  $fd$  values ( $\Sigma fd$ ); sums of the  $fd^2$  values ( $\Sigma fd^2$ ); means of the  $fd$  values or corrections ( $\Sigma fd/N$ , or  $c$ ); means of the  $fd^2$  values ( $\Sigma fd^2/N$ ); and squares of the corrections ( $c^2$ ). Then obtain the standard deviation on each axis in class-interval form ( $\sigma_x$  and  $\sigma_y$ ), using the procedure of steps 3 to 7a, page 337, for computing the standard deviation. As the standard deviations here are to be left in class-interval form, the final operation of multiplying by the size of the class interval (step 7b, page 337) should be omitted.
- (5) Compute the sum of the product moments. Determine the moment (product of  $X$ -axis and  $Y$ -axis deviations) for each cell in which at

least one tally mark appears, obtain the product moment for each such cell by multiplying its moment by its frequency, carry the sum of the product moments in each row to the  $xy$  column, and obtain the sum of the  $xy$  values ( $\Sigma xy$ ).

- (6) Obtain the mean of the product moments. Divide the result of step (5) above by the number of cases ( $\Sigma xy/N$ ).
- (7) Obtain the product of the corrections. Obtain the product of the  $X$ -axis and  $Y$ -axis corrections determined in step (4) above ( $c_x c_y$ ).
- (8) Obtain the product of the standard deviations. Obtain the product of the  $X$ -axis and  $Y$ -axis standard deviations determined in step (4) above ( $\sigma_x \sigma_y$ ).
- (9) Substitute in the formula and solve for  $r$ . Substitute from above the  $\Sigma xy/N$  of step (6), the  $c_x c_y$  of step (7), and the  $\sigma_x \sigma_y$  of step (8) in the formula

$$r = \frac{\frac{\Sigma xy}{N} - c_x c_y}{\sigma_x \sigma_y}$$

and solve to obtain the correlation coefficient.

### 3 MEANING OF CORRELATION COEFFICIENTS

The method of calculating the correlation coefficient as outlined in the foregoing pages is quite mechanical and, as such, can be mastered readily by most students. The interpretation of the meaning or significance of a correlation coefficient is often quite another matter, for no entirely satisfactory mechanical device for accomplishing this has thus far been developed. A number of suggestions have been made recently, however, by means of which the student may be aided in attaching meaning to the correlation coefficient.

#### Not a measure of cause and effect

Sometimes the method of correlation is mistakenly used in the attempt to discover causes operating to produce certain effects. There is nothing in the method or the result of computing a correlation coefficient that indicates definitely which of the factors is a cause and which is an effect or whether both of the factors may be affected by other variables. For example, reading speed and reading comprehension are positively related for pupils in several adjacent school



grades. But it may not be inferred that a pupil reads with high (or low) comprehension *because* he reads rapidly (or slowly). Neither can high (or low) comprehension be thought of as *causing* rapid (or slow) reading. It is likely that some other variable not considered in the correlational relationship, e.g., mental maturity, which is positively related both to reading speed and reading comprehension, serves to explain the relationship. Thus, the *cause* or the explanation of why a positive relationship exists is not necessarily apparent in the data themselves but must often be sought elsewhere.

### Significance for prediction

It will be remembered that correlation is usually indicated by means of what is called a *double-entry table*, *correlation table*, or *scatter diagram*. The appearance of the scatter diagram itself gives some indication of the amount of relationship that exists between the two variables shown. Assuming that the scatter diagram is made by tabulating upward and toward the right, which is almost a universal practice, a high  $r$  is usually found where there is a very definite clustering of the cases along what would be the lower left to upper right diagonal of the table. This means that the cases tend to be grouped somewhat systematically along a line running from the lower left-hand corner of the table to the upper right-hand corner of the table. This type of grouping is shown in Table 28 on page 376. As the cases scatter from the line of this diagonal, the correlation is reduced. If the cases are scattered over the table in a generally circular arrangement, the resulting correlation will approximate zero. As the relationship changes from positive to negative, the elliptical grouping of the cases takes place along a diagonal running from upper left to lower right. After some experience with "scatter diagrams," the student will come to have a definite feeling about what probable magnitude of the correlation coefficient to expect.

One of the important outcomes of the use of correlation methods is that within certain limits it makes possible the estimating of unknown values from known values. The accuracy of this estimate, however, depends directly on the correlation between the factors measured. For example, if it is known from previous experience that there is a high positive relationship between achievement in a specific subject and the pupils' scores on a certain aptitude test, the

probable achievement of a group of pupils in this course may be determined within limits by securing their scores on the aptitude test. A correlation coefficient of  $+1.00$  for the two factors would mean that an estimate of accomplishment based on the one factor would be 100 per cent correct. As the amount of the correlation decreases, the accuracy of the forecast declines, but not in a direct manner. A correlation of  $+1.00$  means 100 per cent accuracy in the estimate based on the relationship, but a correlation of  $+ .50$  does not mean that the estimate based on it will be 50 per cent correct. A glance at the accompanying table will demonstrate this interesting fact about the correlation coefficient. The percentages of forecasting accuracy for different values of  $r$  given in Table 30 are obtained by applying the formula for the *coefficient of alienation* ( $k = \sqrt{1 - r^2}$ ) and then deducting the resulting values, expressed as percentages, from 100. In cases where estimates of one variable are to be made from measurements of another related variable, this table will prove to be a useful safeguard.

TABLE 30. Percentages of forecasting accuracy for certain values of  $r$

Coefficient of Correlation	Percent of Forecasting Efficiency
1.00	100
.99	86
.98	80
.95	69
.90	56
.866	50
.80	40
.75	34
.70	29
.65	24
.60	20
.50	13
.40	8
.30	5
.20	2
.10	$\frac{1}{2}$



A word of warning should possibly be given here in order that the student may not become overoptimistic in the interpretation of correlations. It is better to be on the safe side when making claims for the reliability or the forecasting power of a test. Much damage has been done to the cause of educational measurements by the unqualified and exaggerated statements of misinformed individuals. As a result of frequent misinterpretation of the measures of relationship, many tests of questionable validity and reliability have been accepted and used widely.

Value of $r$	Educational Situation	Interpretation
+ .96	Relation between scores on two forms of a long, analytical reading test for high-school pupils.	Evidence of unusually high reliability; scores may be treated with confidence.
+ .90	Relation between scores on two forms of a 45-minute group intelligence test.	Evidence of marked reliability.
+ .80	Relation between scores on the same form of a group test of intelligence at the beginning and end of a semester.	Evidence of marked relationship; considerable prognostic power even after lapse of time.
+ .50	Relation between scores on a good group intelligence test and course marks of a class in first-year algebra.	Evidence of a medium relationship of little value for forecasting purposes (only 13% effective).
— .24	Relation between chronological ages of pupils in a given grade and scores on an objective achievement test.	Evidence of an indifferent negative relationship; shows a slight tendency for the younger pupils in a grade to achieve at a higher level than the average.

The preceding illustrations and practical interpretations of typical correlation coefficients representative of the sort obtained from educational data have been gleaned from a number of sources. They are offered here for whatever guidance they may give to the student or the teacher in making a critical and conservative interpretation of correlation data.

## 4 PRACTICAL USES OF CORRELATION COEFFICIENTS

The classroom teacher and the student of measurement will find the greatest opportunity to use correlation techniques in connection with the construction and analysis of objective tests and in the critical selection of standardized tests. The uses briefly mentioned and in three cases illustrated below all relate to the determination of test validity, reliability, or objectivity.

### Determination of test validity

Test validity can be determined in terms of correlations between scores on the test and: (1) teachers' marks, (2) ratings of expert judges, (3) other known measures, and (4) measures of future outcomes. All of these situations involve only the ordinary application of the correlation method; therefore, as their values are discussed in Chapter 4, they are not discussed further here.

### Evaluation of test reliability

The correlation coefficient enters directly into the procedures most common for determining or estimating the reliability or consistency of a test, or the degree to which it measures whatever it does measure. As is pointed out more in detail in Chapter 4, there are three correlation methods and two non-correlation methods that can effectively be used by the teacher in estimating the reliability of his classroom tests. These are: (1) the reliability coefficient, (2) the retesting coefficient, (3) the "chance-half" coefficient, (4) the "foot-rule" coefficient, and (5) the standard error of measurement.

*Reliability coefficient.* This coefficient requires only brief mention here, because it involves the usual type of correlational relationship between two series of scores. The reliability coefficient itself is obtained only by correlating scores made by the same pupils on two equivalent forms of the same test.

*Retesting coefficient.* The retesting coefficient, requiring correlation of scores obtained from a first and a second administration of the same test to a group of pupils, furnishes only an estimate of test reliability. The retesting coefficient is one of the methods used when the availability of only one form of the test eliminates the possibility of obtaining a reliability coefficient directly.



"Chance-half" coefficient. This is a second method of estimating the reliability coefficient from the results of the administration of a single test to a pupil group. For this method the first step of procedure is to obtain two "half-scores" for each pupil on arbitrary halves of the test. The arbitrary halves of the test frequently consist of the odd-numbered and the even-numbered items. The second step is to obtain the coefficient of correlation between the sets of half-scores for the group of pupils. This coefficient represents the reliability of *one-half* of the test, but not of the entire test.

The third and final step requires the use of the *Spearman-Brown Prophecy Formula* in estimating the reliability for the entire test by what is known as "stepping up" the correlation. A test increases in reliability as it is increased in length by additional test items comparable to those in the initial test; thus, the estimated reliability for the entire test is greater than for only half of the test. However, the increase in the coefficient is not directly proportional to the increase in test length. The Spearman-Brown formula is

$$r_{12} = \frac{2r_{\frac{1}{2}}}{1 + r_{\frac{1}{2}}},$$

where  $r_{\frac{1}{2}}$  is the correlation between scores on the "chance-halves" of the test and  $r_{12}$  is the estimated reliability coefficient for the entire test.<sup>3</sup>

If an estimate of the reliability of an entire test is desired when the correlation coefficient between its "chance-halves" is .85, the following result is obtained.

$$r_{12} = \frac{2r_{\frac{1}{2}}}{1 + r_{\frac{1}{2}}} = \frac{2 \times .85}{1 + .85} = \frac{1.70}{1.85} = .92.$$

This is the procedure a teacher may use to obtain an estimate of the reliability of his test from a single administration.

<sup>3</sup> The general form of the formula, which is not of direct concern here, is:

$$r_n = \frac{nr_{12}}{1 + (n-1)r_{12}}$$

in which  $r_{12}$  represents the coefficient of reliability of a test and  $r_n$  represents the coefficient of reliability of a test of homogeneous test materials  $n$  times as long. It should be noted that substitution of 2 for  $n$  in this formula, to determine the effect of doubling the length of the test, results in the special formula given above except for differences in the subscript for  $r$ .

"Footrule" coefficient. This coefficient is a third and quite simple method of obtaining an estimate of the reliability of a test available in only one form.<sup>4</sup> The only values required are the arithmetic mean and standard deviation of the test scores and the number of items in the test. The formula is

$$r_{tt} = \frac{n}{n-1} \times \frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2},$$

where  $\bar{p} = \frac{M_t}{n}$  and  $\bar{q} = 1.00 - \bar{p}$ , and where  $M$  is the arithmetic mean of the test scores,  $\sigma_t$  is the standard deviation of the test scores, and  $n$  is the number of test items.

The "Footrule" coefficient of a test of 249 items for which the arithmetic mean and standard deviation of scores were respectively 168.65 and 25.34 would be obtained by the following procedures

$$\begin{aligned}\bar{p} &= \frac{M_t}{n} = \frac{168.65}{249} = .677 & \bar{q} &= 1.00 - \bar{p} = .323 \\ r_{tt} &= \frac{n}{n-1} \times \frac{\sigma_t^2 - n\bar{p}\bar{q}}{\sigma_t^2} \\ &= \frac{249}{248} \times \frac{25.34^2 - 249 \times .677 \times .323}{25.34^2} \\ &= 1.004 \frac{642.116 - 54.45}{642.116} = 1.004 \times .915 = .919\end{aligned}$$

*Standard error of measurement.* This measure of the confidence that may be placed in an obtained score is based on the reliability coefficient of the test, but it supplies a more concrete and invariable measure of consistency than does the reliability coefficient itself. This is a result of the fact that the standard error of measurement is not, as is the reliability coefficient, influenced by varying ranges of talent in the pupil group upon which the measure is based.

Computation of the standard error of measurement involves the use of the formula  $S.E._m = S.D. \sqrt{1 - r}$ , where  $S.E._m$  is the standard error of measurement,  $S.D.$  is the standard deviation of the distribution of scores, and  $r$  is the reliability coefficient of the test. The standard deviation of the distribution of 37 reading test scores used

<sup>4</sup> G. F. Kuder and M. W. Richardson, "The Theory of the Estimation of Test Reliability," *Psychometrika*, 2:151-60; September 1937 (Formula 21).



to illustrate computational procedures in the preceding chapters was shown in Table 17, page 333, to be 10.86. The reliability coefficient of the test, based on the correlation coefficient between the scores reported in Table 9 and scores made by the same 37 pupils on a comparable form of the test, was found to be .956. Using this reliability coefficient and the value of the *S.D.* shown above in the formula, the standard error of measurement becomes

$$\begin{aligned} S.E._m &= 10.86 \sqrt{1 - .956} \\ &= 10.86 \sqrt{.044} \\ &= 10.86 \times .210 \\ &= 2.28 \end{aligned}$$

Although a complete explanation of how to apply the standard error of measurement is not feasible here, its general use can be illustrated. It will be remembered that a pupil's obtained score is only an estimate of what his true score would be if the test were completely reliable. As no test is absolutely reliable, an obtained score must be interpreted as an estimate of the true score. The standard error of measurement indicates how far the obtained score may be expected to deviate from the true score. An illustration will be given in terms of a score of 48 made by a certain pupil on this test. His obtained score is almost certain to fall within three standard errors of his true score. Therefore  $3 \times 2.28$ , or 6.84, is subtracted from and added to 48 ( $48.00 - 6.84 = 41.16$  and  $48.00 + 6.84 = 54.84$ ). Thus it is a practical certainty that the pupil's true score on this test lies between 41.16 and 54.84. Again, the chances are about two in three that his obtained score lies within one standard error of his true score. Thus  $48.00 - 2.28$  and  $48.00 + 2.28$ , or 45.72 and 50.28, indicate the limits of his obtained score from his true score that are not likely to be exceeded more than once in three times.

Although there are no definite standards for applying this type of check on the reliability of a test score, it may be said that in general a test is more reliable when the standard error is small than when it is large. It is apparent from the above formula that the standard error is small when test reliability is high and large when test reliability is low. This discussion of the reliability of test scores should illustrate the fact that too great dependence on test scores is unwarranted, for a test score is at best only an estimate, and an

estimate subject to some degree of error, of what the corresponding true score would be if it were obtainable.

### Determination of test objectivity

When a group of test papers has independently been scored twice, either by the same person or by different persons, the correlation coefficient between the two sets of scores is the objectivity coefficient. For a highly objective test, the coefficient should closely approach  $+1.00$ .

Pupil	Total Scores		Vocabulary Scores		Pupil	Total Scores		Vocabulary Scores	
	Read.	Vocab.	Odd	Even		Read.	Vocab.	Odd	Even
a	65	101	52	49	u	64	82	43	39
b	83	106	54	52	v	55	74	40	34
c	52	83	43	40	w	58	67	36	31
d	74	97	47	50	x	69	75	39	36
e	48	54	27	27	y	84	117	59	58
f	57	49	26	23	z	66	85	44	41
g	75	86	45	41	aa	69	92	47	45
h	60	103	51	52	ab	73	86	46	40
i	81	113	57	56	ac	71	79	40	39
j	70	96	50	46	ad	62	91	47	44
k	78	94	50	44	ae	57	69	33	36
l	54	88	42	46	af	60	71	37	34
m	59	89	41	48	ag	67	81	45	36
n	71	106	51	55	ah	50	63	30	33
o	50	46	24	22	ai	66	86	42	44
p	45	49	26	23	aj	54	92	47	45
q	61	70	35	35	ak	67	74	36	38
r	90	114	59	55	al	62	78	41	37
s	78	101	50	51	am	86	97	49	48
t	53	79	37	42	an	68	74	35	39

### Problems on Computing the Correlation Coefficient and Estimating the Reliability Coefficient

27. Compute the Pearson product-moment correlation coefficient between the reading and vocabulary test scores listed above for 40 ninth-grade pupils. ( $r = +.731$ )



28. Obtain an estimate of the reliability coefficient of the vocabulary test by computing the Pearson product-moment correlation coefficient between the "odd" and "even" scores listed above for 40 ninth-grade pupils and then using the Spearman-Brown Prophecy Formula. ( $r_{12} = .954$ )
29. Compute the "Footrule" coefficient for a history test consisting of 120 items on which a class of ninth-grade pupils attained an arithmetic mean and a standard deviation of 71.20 and 12.60 respectively. ( $r_{tt} = .819$ )
30. Compute the standard error of measurement for the vocabulary test of Problem 27, using the  $r_{12}$  of .954 and the standard deviation, in score form, of 12.40. ( $S.E._m = 2.73$ )

### Selected References

- FROELICH, CLIFFORD P., AND DARLEY, JOHN G. *Studying Students: Guidance Methods of Individual Analysis*. Chicago: Science Research Associates, 1952. Chapters 3, 12.
- GARRETT, HENRY E. *Statistics in Psychology and Education*. Fourth edition. New York: Longmans, Green and Co., 1953. Chapter 6. p. 332-44.
- GERBERICH, J. RAYMOND, AND PETERS, CHARLES C. "Reliability." *Encyclopedia of Modern Education*. New York: Philosophical Library, 1943. p. 673-75.
- GREENE, HARRY A., AND CRAWFORD, JOHN R. *Work-Book in Educational Measurements and Evaluation*. New York: Longmans, Green and Co., 1945. Unit 10.
- LINDQUIST, E. F. *A First Course in Statistics*. Revised edition. Boston: Houghton Mifflin Co., 1942. Chapters 10-11.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. p. 94-101.
- ODELL, C. W. *An Introduction to Educational Statistics*. New York: Prentice-Hall, Inc., 1946. Chapters 7, 9.
- REMMERS, H. H., AND GAGE, N. L. *Educational Measurement and Evaluation*. New York: Harper and Brothers, 1943. p. 537-50.
- ROSS, C. C. *Measurement in Today's Schools*. Second edition. New York: Prentice-Hall, Inc., 1947. p. 237-50.
- WALKER, HELEN M. *Elementary Statistical Methods*. New York: Henry Holt and Co., 1943. Chapter 12; p. 222-35.

- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. p. 59-69.
- WEITZMAN, ELLIS, AND MCNAMARA, WALTER J. *Constructing Classroom Examinations*. Chicago: Science Research Associates, 1949. Chapter 11.



## *Measuring and Evaluating in the Receptive Language Arts*

THE FOLLOWING important points involved in the measurement and evaluation of listening and reading skills are summarized in this chapter:

- A. The educational and social significance of listening and reading.
- B. Major objectives of listening and reading.
- C. Testing reading readiness.
- D. Testing listening and oral reading.
- E. Testing and remediation in work-type reading.
- F. Types of remedial material in reading.

The receptive language arts, which are here considered to consist of listening, reading, and the study skills, are distinguished from the expressive language arts, which include oral and written language, usage, grammar, spelling, and handwriting. Both the receptive and the expressive aspects of language are included in the basic skills stressed in the elementary school. They are also often designated as communication skills. The receptive or assimilative language arts are dealt with in this chapter, while the following chapter is concerned with the expressive or outgoing forms of language. Treatments in both chapters are confined to the English language because of the fact that the foreign languages are seldom taught in American elementary schools.

## I IMPORTANCE OF LISTENING AND READING AS RECEPTIVE LANGUAGE SKILLS

### Educational and social significance of listening and reading

There is a growing conviction on the part of students of the language arts areas that, of the two language channels through which information is received, listening has been seriously neglected in favor of reading. Reading has been analyzed, investigated, and evaluated perhaps more than any other school subject. The educational literature of the past decade is heavily loaded with articles, books, devices, and tests dealing with reading. Listening, on the other hand, appears relatively infrequently in the literature. As Brown <sup>1</sup> pointed out "if the average individual depended less upon listening than upon reading, there might be reason for this neglect." Studies and observations dating back over a twenty-year period indicate that listening is without question one of the most frequently used language activities. Actually, the average adult spends approximately three times as much time in listening as he does in reading.<sup>2</sup> If this is the case, there is ample reason for the growing belief that listening skills should be developed as a part of a systematic program of instruction in the language arts.

In a survey designed to discover (1) what percentage of the school day children are expected to listen, (2) whether teachers themselves are aware of the amount of listening children are expected to do, and (3) what relative importance teachers place upon the four phases of language education, Wilt <sup>3</sup> concluded, among other things, that children are expected to spend more time in listening than in any other single activity of the school. Teachers apparently are unaware of the above fact. Moreover, they apparently are more concerned about the child who is reading aloud or speaking than they are about the listeners. Almost fifteen hundred teachers estimated that children learn through reading approximately 110 minutes of the average school day, and through listening 78 minutes. Actually, the median

<sup>1</sup> James I. Brown, "The Construction of a Diagnostic Test of Listening Comprehension." *Journal of Experimental Education*, 18:139-46; December 1949.

<sup>2</sup> Paul T. Rankin, "Listening Ability: Its Importance, Measurement and Development." *Chicago Schools Journal*, 12:177-79; January 1930.

<sup>3</sup> Miriam E. Wilt, "A Study of Teacher Awareness of Listening as a Factor in Elementary Education." *Journal of Educational Research*, 43:626-36; April 1950.



amount of listening time was 158 minutes, 54 per cent of which was spent in listening to the teacher, 31 per cent in listening to the other children, and 15 per cent in miscellaneous listening. While no comparable data on this problem are available at the high-school or college levels it is quite probable that the relative importance of listening as a factor in learning at these levels is even more significant than at the elementary-school level. The logical conclusions from this train of thought should result in making all teachers sensitive to the importance of listening as a factor in intelligent communication and in practically all types of learning. This in turn should result in granting to listening a place of importance in the curriculum at least equal to that of reading.

The relatively large proportion of classroom time given over to listening may seem somewhat less serious when it is remembered that only a part of learning occurs in the classroom. Even in the elementary grades a large amount of learning through reading takes place. The solution of most classroom problems in the modern school requires the skillful use of books as sources of information. When considered from this point of view, reading as a school responsibility is something more than merely the rapid comprehension of printed symbols and the memory and organization of the materials read. It is also the ability to utilize books and libraries as efficient sources of information. This tendency to treat reading as a highly important tool of learning has resulted in establishing a very close relationship between reading and practically every other school activity. As a means of gaining information and pleasure it is essential in every content subject, such as history, geography, science, literature, and arithmetic.

A full appreciation of the importance of intelligent reading in society at large has also developed in recent years. Reading is considered the indispensable means by which adults may keep abreast of current happenings and familiarize themselves with current social, community, political, and national problems. The mass of printed matter which the typical adult must read and evaluate, even within the limits of his own fields of interest, is stupendous. This situation makes all the more imperative the development of a high degree of reading skill in our schools.

## Major objectives and outcomes in reading and listening

The necessity for the development of a high level of reading ability along with effective skill in critical and evaluative listening on the part of all children and adults is more readily realized when it is recognized that a large share of the vast bulk of facts, informations, and skills they are supposed to master are obtained through these avenues. The real significance of the matter is seen in the fact that the level of reading and listening ability on the part of many of these individuals is not particularly high.

The extremely wide variety of school and life situations in which children and adults read or listen is indicated in the following list of reading and listening objectives, attitudes, and abilities. The outline itself is an adaptation of material presented by Greene and Gray<sup>4</sup> in a discussion of the measurement of understanding in the language arts.

### A FUNCTIONAL ANALYSIS OF READING AND LISTENING

#### I. Objectives in Reading and Listening

##### A. Typical life situations which lead children and adults to read or listen

1. To find out what is going on
2. To find one's way about
3. To understand directions and assignments
4. To verify spellings, pronunciations, meanings, use of words
5. To secure answers to specific questions
6. To gather information for fuller understanding, or for informing or convincing others
7. To learn how to act in new situations
8. To work out complicated problems
9. To reach conclusions as to guiding principles, relative values, or cause-effect relationships
10. To identify and resolve propaganda
11. To search for and discover the truth

<sup>4</sup> Harry A. Greene and William S. Gray, "The Measurement of Understanding in the Language Arts." *The Measurement of Understanding*, Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. University of Chicago Press, Chicago, 1946. p. 189-200. Quoted by permission of the Society.



- B. Typical recreational situations which lead children and adults to read or listen
  - 1. To relive everyday experiences
  - 2. To have fun or sheer enjoyment
  - 3. To escape from real life
  - 4. To satisfy curiosities about strange times and places, human nature, and motives
  - 5. To enjoy sensory imagery
  - 6. To enjoy ready-made emotional reactions through hearing or reading romantic tales, sentimental verses, mystery stories
  - 7. To enjoy the sentiments and ideals expressed by others
  - 8. To enjoy the rhythm and quality of expression in both prose and poetry
- II. Basic Reading and Listening Knowledges, Attitudes, and Skills
  - A. Responding to the motive, problem, or purpose of a statement
  - B. Directing attention to the meaning of what is read or heard
  - C. Developing fluent, accurate perception of word forms
  - D. Recognizing and using new words and meanings
  - E. Securing an adequate understanding of what is read or heard
    - 1. To grasp meanings of words appropriate to the context
    - 2. To fuse word meanings into a chain of related ideas
    - 3. To recognize the relationship and importance of ideas
    - 4. To handle unusual word order, complex sentence structure, abstract ideas
    - 5. To interpret meaning in the light of the total setting, the author's or the speaker's tone and intention
    - 6. To supplement the specific meanings by reading between the lines, drawing inferences, seeing implications
  - F. Reacting critically to what is heard or read
    - 1. To realize the significance of the ideas presented
    - 2. To judge the validity of the ideas presented
    - 3. To evaluate the soundness, accuracy, or completeness of the author's or speaker's conclusions, and the accuracy of his reasoning
  - G. Blending the ideas acquired with previous experience
    - 1. To acquire new insights
    - 2. To reaffirm or modify previous understandings
    - 3. To solve critical problems
    - 4. To acquire rational attitudes
    - 5. To modify behavior
    - 6. To broaden interest

### III. Attitudes, Skills, and Procedures Essential in Work-Study Type Reading

- A. Comprehending quickly what is read
  - 1. To use rapid and rhythmic eye-movements
  - 2. To avoid lip reading
  - 3. To correctly associate symbols, words, and meanings
- B. Locating needed information
  - 1. To understand and use an index
  - 2. To use a table of contents
  - 3. To use the dictionary
  - 4. To use library card files
  - 5. To use reference books
  - 6. To use and interpret maps, graphs, and tables
- C. Gathering and evaluating information in the light of a given purpose
  - 1. To recognize the purposes to be achieved, by
    - a. Finding answers to specific questions
    - b. Finding the central thought of the selection
    - c. Following a sequence of related events
    - d. Enjoying the facts or the story presented
    - e. Identifying important points and supporting details
    - f. Selecting facts relating to the problem
    - g. Solving a specific problem
    - h. Understanding and following directions
    - i. Comparing the views of authorities
    - j. Supporting a point of view or a course of action
  - 2. To apply appropriate fact-finding techniques such as
    - a. Studying the title for cue to its meaning
    - b. Reading carefully to discover what the author plans to do or say
    - c. Noting especially topic sentences or paragraphs
    - d. Noting the author's method of arriving at his point
    - e. Grasping the author's organization of ideas
  - 3. To separate essential from non-essential information
  - 4. To judge the significance of relevant information
  - 5. To organize information in terms of the specific problem
    - a. Summarizing
    - b. Outlining
  - 6. To draw tentative conclusions defensible in the light of the facts
  - 7. To decide when the purpose has been achieved
  - 8. To give credit to sources of facts and information



- D. Adjusting reading attitudes and procedures to different purposes
  - 1. To select and remember relevant facts in reading to answer factual questions
  - 2. To note the author's organization of facts, select, associate, remember, and reorganize them in preparing a report
  - 3. To select relevant facts, compare them with other known facts, and judge their validity in determining the accuracy of facts or events described
    - a. Memorizing quickly
    - b. Reporting without notes
  - 4. To read slowly and carefully when a thorough understanding of relatively difficult material is involved
  - 5. To read rapidly when the purpose is to find out what is in the article or to enjoy a story
  - 6. To skim rapidly when hunting for relevant material or locating specific items of information
- IV. Attitudes, Skills, and Procedures Essential in Interpretative Oral Reading
  - A. Insuring a thorough grasp of the author's meaning by utilizing those skills specified for the purpose in the foregoing outline
  - B. Developing a clear, pleasant, properly modulated voice, clear enunciation of words, and correct pronunciation of words
  - C. Having a compelling motive for reading to others
  - D. Sensing the importance of the message for the listening audience
  - E. Adjusting manner and speaking voice to the size of the room, character of the selection, and needs of the audience
  - F. Modulating voice to bring out thought relationships clearly
  - G. Adjusting the voice to changes in character and mood
  - H. Adjusting rates of reading and the grouping of words to the rhythm of poetry
  - I. Using appropriate facial expression and gesture, subordinated to the thought of the selection
  - J. Controlling breathing and bodily movements
  - K. Feeling confident of ability, free from tension, natural, sincere, and convincing in manner and speech

The foregoing outline of the major objectives of listening and reading affords a useful basis for the evaluation of present instructional emphasis in these two important acquisitive skill areas as well as a valuable source of criteria for the validation of analytical and corrective procedures in listening and reading.

## 2 IDENTIFICATION OF FACTORS AFFECTING LISTENING AND READING

### Factors in listening efficiency

It seems safe to assume that listening, like reading, is a composite of several somewhat independent but related skills. In addition to intelligence and reading comprehension, Nichols<sup>5</sup> indicated that such factors as the following appear to influence the individual's listening comprehension:

1. Recognition of correct English usage
2. Size of the listener's vocabulary
3. Ability to make inferences
4. Ability to see the organization plan of a speech
5. Ability to listen for main ideas rather than for specific facts
6. Use of special techniques for the improvement of concentration
7. Real interest in the subject discussed
8. Physical fatigue
9. Audibility of the speaker
10. Respect for listening as a means of learning
11. Susceptibility to distractions
12. Experience in listening

In an attempt to identify the basic measurable factors in listening comprehension, Brown<sup>6</sup> concluded that the following skills are involved in the effective use of listening as a learning instrument:

1. Identification and recall of details presented orally
2. Ability to follow the sequence of details in the form of oral directions
3. Retention of details long enough to answer questions about them
4. Ability to listen reflectively for the purpose of identifying the central idea of the statement given orally
5. Ability to draw inferences from the supporting facts presented in the statement
6. Ability to distinguish relevant from irrelevant materials
7. Use of contextual clues to word meanings
8. Recognition of transitional elements in sentences

On the basis of these factors Brown has proposed an analytical test of listening comprehension.

<sup>5</sup> Ralph G. Nichols, "Factors in Listening Comprehension." *Speech Monographs*, 15:154-63; Research Annual, No. 2, 1948.

<sup>6</sup> Brown, *op. cit.* p. 140-41.



## Typical defects in reading

The solution of the problem of the effective initial teaching of reading as well as the development of satisfactory remedial materials in reading is dependent to a large degree on the accurate identification of the specific causes of reading failure. Not only is it necessary to discover the child who in his later school experience is almost certain to encounter reading difficulties, but these reading difficulties must be identified definitely and accurately. Harris listed and discussed at length <sup>7</sup> the following causes of reading difficulties: (1) low intelligence, (2) visual defects, (3) auditory defects, (4) other physical conditions—defects of muscular coordination and speech, glandular disturbances, and neurological difficulties, (5) lack of hemispherical dominance, (6) poor school record, (7) deficiencies in arithmetic, spelling, and handwriting, and (8) emotional and social problems. He pointed out, <sup>8</sup> however, that it is “impossible to determine the relative contribution of each handicap to the total picture of failure. . . . From a practical standpoint, the aim of a thorough diagnosis is not to fix the blame for the child’s difficulties, but to discover each of the many conditions that may require correction.”

## Oral vs. silent reading

An examination of the major aspects of remedial work in reading indicates that there are two angles from which it may be considered. In the first place, remedial instruction may be begun in the oral reading field. Gray and others have defended this point of attack on the problem on the ground that it enables the teacher to start with the child on a level at which he already has some mastery, that is, the oral language level. Others believe that on account of the large proportion of reading time spent in the work-type of silent reading this field should receive the special emphasis. There is merit on both sides of the question, undoubtedly. It is true that the child does come to school with a fairly adequate oral vocabulary, which in a great many ways affords the natural approach to reading. On the other hand, it is also true that such an approach tends to place too large

<sup>7</sup> Albert J. Harris, *How to Increase Reading Ability: A Guide to Individualized and Remedial Methods*, Second edition. Longmans, Green and Co., New York, 1947. Chapter 7.

<sup>8</sup> *Ibid.* p. 242.

an emphasis on the pronunciation of words and too little on their meaning when encountered in silent reading situations. The transfer from the emphasis on oral reading (pronunciation of words) to silent reading (comprehension of meaning of words, sentences, and paragraphs) must be made at some point in the child's experience. Accordingly, a great many teachers hold that the place to start the emphasis on silent reading is at the beginning. Some foundation for this belief is seen in the results obtained by many teachers who place the emphasis on the development of silent reading skills at the outset.

## Readability

During the past decade teachers have shown a surprising increase of interest in the objective, valid, and reliable evaluation of the suitability of textbooks and other reading material for classroom use. One aspect of this type of textbook evaluation involves readability, or the understandability of printed material. Betts<sup>9</sup> pointed out that current interest appears to be centered on the language and the content of reading material. Recent workers have concerned themselves with relationships between these factors in readability: vocabulary difficulty, vocabulary diversity, sentence length or structure, "human interest," and meaning. On the basis of these factors, formulae have been derived for predicting the difficulty of reading material. Objective measures of readability are given precedence over author and teacher judgments.

The early work of Vogel and Washburne in preparing the *Winnetka Graded Book List* provided the basis for the concept of readability as well as for the general method of measuring it. In their first formula they weighted four factors: (1) number of different words, (2) number of uncommon words, (3) number of prepositions, and (4) the number of simple sentences.<sup>10</sup> Ten years later they reported a revised readability formula based upon only three elements: number of different words, number of uncommon words, and the number of simple sentences.<sup>11</sup> According to Lorge "the pattern established by this formula has been followed by Lewerenz (1929), Ojemann (1933),

<sup>9</sup> Emmett A. Betts, "Readability: Its Application to the Elementary School." *Journal of Educational Research*, 42:438-59; February 1949.

<sup>10</sup> Mabel Vogel and Carleton Washburne, "An Objective Method of Determining Grade Placement of Children's Reading Material." *Elementary School Journal*, 28:373-81; January 1928.

<sup>11</sup> Carleton Washburne and Mabel Vogel, "Grade Placement of Children's Books." *Elementary School Journal*, 38:355-64; January 1938.



Dale and Tyler (1934), Gray and Leary (1935), Lorge (1939), Flesch (1943), and by Dale and Chall (1948). In each instance a multiple regression formula was developed relating a criterion and some internal indications of expressional difficulty."<sup>12</sup> Four kinds of elements appear to have been considered in most of these readability formulae: (1) vocabulary load, (2) sentence structure, (3) idea density, and (4) human interest. Yoakam, in an early unpublished study, indicated that vocabulary load was a sufficiently reliable index of readability. Lorge<sup>13</sup> concluded that a weighted index of vocabulary load is one of the best measures of difficulty in texts planned for use below the fourth or fifth grades.

Readability formulae if used within proper limits make available to the classroom teacher very valuable means of evaluating written materials presented to the child. It should be understood that thus far they reflect only certain mechanical aspects of understandability. They do not necessarily reveal anything concerning the quality, complexity, or relationships of the ideas expressed.

### 3 DETERMINATION OF READING READINESS

#### Factors in reading readiness

Reading readiness is dependent on a large number of characteristics. Harris listed the following as among the most important: (1) intelligence, (2) visual perception, (3) auditory perception, (4) language development, (5) background of experience, and (6) social behavior.<sup>14</sup> It is unsafe to assume that a child who enters school at the age of six is ready for reading. Some children have already learned to read, and are mentally much more mature than the average child of six, while others may have no more mental maturity than the average child of four, and may thus encounter considerable difficulty in learning to read.

Gates<sup>15</sup> recommended that "before the child actually begins to learn to read his status should be determined in the following respects: (1) intelligence or verbal aptitude, (2) vision, (3) color blindness,

<sup>12</sup> Irving Lorge, "Readability Formulae—An Evaluation." *Elementary English*, 26:86-95; February 1949.

<sup>13</sup> *Ibid.* p. 92.

<sup>14</sup> Harris, *op. cit.* p. 48.

<sup>15</sup> Arthur I. Gates, *The Improvement of Reading: A Program of Diagnostic and Remedial Methods*, Third edition. Macmillan Co., New York, 1947. p. 142.

(4) hearing, (5) handedness, (6) speech, (7) health and vigor, and (8) emotional stability."

Betts<sup>16</sup> summarized his observations on the experimental evidence on the problem of reading readiness as follows:

Reading is a very complex process, no one factor stands out in bold relief. . . . Factors in reading readiness are inextricably interrelated. Furthermore, each factor carries a different weight in predicting readiness for reading. The teacher deals with the total organism of a growing child. Factors (in reading readiness) are highly significant at all levels and in all areas of instruction. Each teacher is a first-hand dealer in systematic sequences of readiness. . . . It is especially important to keep these factors in mind when developing a differentiated reading-readiness program. These factors are the ingredients of a compound called "reading readiness."

Reporting on the basis of his experiences in the reading clinic, Betts<sup>17</sup> made the following significant statement:

It is imperative that a teacher should not drive a child into reading until she has made an attempt to analyze or define the problem. Our records show that almost 90 per cent of the severe reading cases should have medical attention before receiving pedagogical help. In such instances, tutoring aggravates the problem and many times an apparent gain in reading achievement is due to maturation rather than to the pedagogical methods used.

Such cases depend for their effective remediation on the services of persons other than the teacher, obviously, but they are no less a classroom concern for that reason.

## Reading readiness tests

Reading readiness tests can best be classified as tests of specific intelligence, for their purpose is to measure the mental ability factors essential to success in reading. These factors are measured by tests that make use of visual and auditory abilities that are basic to reading. Reading readiness tests employ several testing devices. Among them are: (1) distinguishing pictured objects that are named

<sup>16</sup> Emmett A. Betts, *Foundations of Reading Instruction with Emphasis on Differential Guidance*. American Book Co., New York, 1946. p. 137-38.

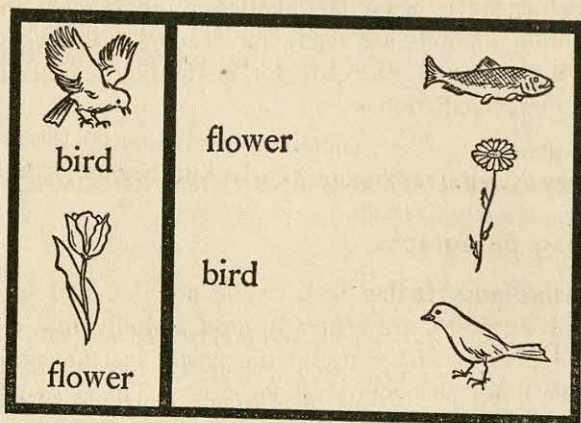
<sup>17</sup> Emmett A. Betts, "Teacher Analysis of Reading Disabilities." *Elementary English Review*, 11:99-102; April 1934.



by the examiner, (2) matching one word of a group with its counterpart, which appears as a visual stimulus, (3) recognizing word similarities or differences, (4) recognizing rhyming words, and (5) reading numbers and letters. Only the last of these can perhaps be called a reading skill, although all of the types measure abilities on which later reading abilities depend. The major purpose of reading readiness tests is to locate those children who are not yet ready to start reading but for whom that activity should be delayed until the child's mental maturity and experience are adequate for such an undertaking.

The accompanying illustration taken from the *Harrison-Stroud Reading Readiness Tests* is representative of testing methods for reading readiness. Instructions are given orally for all such tests. The following five tests comprise this series: (1) Making Visual Discriminations, (a) Attention Span Controlled, (b) Attention Span Uncontrolled; (2) Using the Context; (3) Making Auditory Discriminations; (4) Using Context and Auditory Clues; and (5) Using Symbols. The accompanying exercise is from Test 5.

Excerpt from Harrison-Stroud Reading Readiness Tests <sup>18</sup>



The *Gates Reading Readiness Test*, issued to accompany his reading comprehension tests, which will be discussed in the following pages, consists of five parts, the last one of which must be admin-

<sup>18</sup> M. Lucille Harrison and James B. Stroud, *The Harrison-Stroud Reading Readiness Tests*. Published by Houghton Mifflin Co., 1950.

istered to one child at a time. The five parts of the Gates test are: (1) picture directions, (2) word matching, (3) word-card matching, (4) rhyming, and (5) letters and numbers.

Other reading readiness tests embody in general the same measurement devices. The *Lee-Clark Reading Readiness Test* requires pupils to match similar letters, to recognize the one letter that is unlike the other three of a group, and to cross out the extra letter occurring in one of two words otherwise alike. The *Monroe Reading Aptitude Tests* provide five types of tests: (1) visual, for measuring memory for position of visual forms, control of eye movements, and drawing from memory, (2) auditory, for measuring ability to detect correct pronunciations, to distinguish between words that sound somewhat alike, and to reproduce a story from memory, (3) language, for measuring extent and richness of vocabulary, (4) articulation, for measuring correctness of articulation and speed in repeating words, and (5) laterality, for measuring hand, eye, and foot preferences.

The *Metropolitan Readiness*, the *Van Wagenen Reading Readiness*, the *Betts Ready to Read*, and the *Murphy-Durrell Diagnostic Reading Readiness* tests are all useful tests of the pupil's background for reading instruction. The *Durrell-Sullivan Reading Capacity Tests* are other tests which serve the double function of providing an indication of whether or not pupils are ready for reading instruction and of providing for those who are ready for instruction a basis for their classification or remediation.

#### 4 MEASURING ORAL READING AND LISTENING COMPREHENSION

##### Oral reading paragraphs

The available oral reading tests, while not designed specifically for diagnostic purposes, are generally used as individual tests, and accordingly represent rather useful diagnostic instruments for the teacher of the lower elementary-school grades. The *Gray Standardized Oral Reading Paragraphs*, which have been very widely used over a period of years, consist of twelve paragraphs arranged in increasing order of difficulty. The tests are administered by having the pupils read the paragraphs under time control until a certain number of errors per paragraph are made.

The *Gilmore Oral Reading Test*, a much newer device, provides measures of accuracy of oral reading, comprehension of material



read, and rate of reading. Each of the two equivalent forms comprises ten oral paragraphs arranged in ascending order of difficulty. Vocabulary, sentence structure, and interest are carefully controlled as factors in the difficulty of the paragraphs. There are five comprehension exercises for each paragraph.

### Excerpts from Gilmore Oral Reading Test<sup>19</sup>

- 3.** The name of the boy is Bob.  
The name of his sister is Jane.  
They live with their parents  
in a white house near the city.  
They are playing on the walk.  
The dog and cat are their pets.  
After Father has gone to work,  
the children will leave for school.

TIME \_\_\_\_\_Seconds

1. What is the boy's name?
2. What is his sister's name?
3. Where is their house?
4. What are their pets?
5. When will the children leave for school?

NUMBER RIGHT \_\_\_\_\_

ERROR RECORD	Number
Substitutions	
Mispronunciations	
Words pronounced by examiner	
Disregard of punctuation	
Insertions	
Hesitations	
Repetitions	
Omissions	
Total Errors	

- 4.** Mother waves good-by to Father each morning. She begins the housework soon after he leaves. Bob and Jane help her before they go to school. They dry the dishes and clean their own rooms. After Mother has finished the work indoors, she goes out to her pretty flower garden. She tends it nearly every day for about an hour. Mother does all her work with great care.

TIME \_\_\_\_\_Seconds

1. What does Mother do as Father is leaving?
2. What does Mother do after Father has gone?
3. When do Bob and Jane help Mother?
4. Where does Mother go after she has finished the work indoors?
5. How long does she work in her garden each day?

NUMBER RIGHT \_\_\_\_\_

ERROR RECORD	Number
Substitutions	
Mispronunciations	
Words pronounced by examiner	
Disregard of punctuation	
Insertions	
Hesitations	
Repetitions	
Omissions	
Total Errors	

There are ten paragraphs which are arranged in ascending order of difficulty and which form a continuous story related to the same characters. A separately bound booklet contains the illustrations and the paragraphs for the use of the subject being tested. The type of material presented on the record booklet for the use of the examiner is illustrated by the accompanying samples, Paragraphs 3 and 4 from Form A of the Test.

### Oral reading check tests

The plan of recording the number and the kinds of errors made by the pupil in reading the *Gray Standardized Oral Reading Paragraphs* permits a type of diagnostic analysis of oral reading abilities. Much more concise information of this kind is made available, however, through the use of Gray's *Oral Reading Check Tests*.

As in the oral reading paragraphs, these check tests are to be given

<sup>19</sup> John V. Gilmore, *Gilmore Oral Reading Test*. Published by World Book Co., 1952.

individually. The errors made by the pupil are recorded by the teacher on a separate test sheet showing the types of errors made by the pupil which appear most frequently in oral reading. The following illustration may make clear the character of the errors and the method of recording them: <sup>20</sup>

The sun pierced into my <sup>many</sup> large windows. It was the opening of October, and the <sup>clear</sup> sky was of a dāzzling blue. I looked out of my window and down the street. The white houses of the long, straight street were almost painful to the eyes. The clear atmosphere allowed full play to the sun's brightness.

If a word is wholly mispronounced, underline it as in the case of "atmosphere." If a portion of a word is mispronounced, mark appropriately as indicated above: "pierced" pronounced in two syllables, sounding long *a* in "dazzling," omitting the *s* in "houses" or the *al* from "almost," or the *r* in "straight." Omitted words are marked as in the case of "of" and "and"; substitutions as in the case of "many" for "my"; insertions as in the case of "clear"; and repetitions as in the case of "to the sun's." Two or more words should be repeated to count as a repetition.

The individual record sheet that accompanies Gray's *Oral Reading Check Tests* is useful in two important ways. It places before the teacher a carefully classified list of common errors in oral reading, and it provides space for the recording of successive repetitions of the test so that progress may be measured.

The analysis of the individual pupil's record gives a very concise picture of his oral reading difficulties. It will be noted that in these oral reading exercises no attention is paid to the degree of comprehension with which the material is read. The measurement of comprehension lies somewhat beyond the purpose of this test. Here the purpose is the determination of the efficiency with which words are recognized and pronounced in context, with little or no concern for the comprehension of the materials.

## Measuring listening comprehension

Tests for the measurement of the comprehension of orally presented material at the elementary-school level are limited in number

<sup>20</sup> William S. Gray, *Oral Reading Check Tests*. Published by Public School Publishing Co.



and in scope. In fact, no standardized listening comprehension test is available at this time except portions of certain reading readiness tests. The importance of auditory comprehension tests as a part of reading readiness is easily understandable. In the light of the great importance of listening as a learning tool, it is not clear why so little has been done with it at the upper primary and intermediate grade levels.

A lack of auditory discrimination was recognized by Murphy and Durrell<sup>21</sup> as one of the three most important causes of pupil failure in learning to read. The first section of the *Murphy-Durrell Diagnostic Reading Readiness Test* consists of 84 items designed to determine the ability of the pupil to recognize similarities and differences in the sounds of words by comparing the name of a picture and the sound of a word. The first group of items tests the pupil's ability to sense the differences in the beginning sounds of words; the second group tests the ability to sense final sounds.

The *Brown-Carsen Listening Comprehension Test*,<sup>22</sup> in spite of the fact that it is standardized for high-school and college use only, suggests interesting possibilities for similar work at the elementary-school level. The test is composed of five parts: (1) Immediate Recall, (2) Following Directions, (3) Recognition of Transitions, (4) Recognizing Word Meanings, and (5) Lecture Comprehension. The test, of course, is administered orally. Student's responses are recorded on separate answer sheets.

## 5 ANALYSIS AND DIAGNOSIS IN SILENT READING

### Measurement of work-study type of reading

The emphasis given to the work types of reading in the list of skills given on pages 395 to 398 indicates something of the importance of this type of reading in relation to the total reading field. Some pupils fail (in arithmetic, for example), not entirely because of ignorance of the basic facts, or lack of mental ability to understand the explanations, but rather on account of sheer inability to read. In fact, one of the best ways to improve work in many other school

<sup>21</sup> Helen A. Murphy and Donald D. Durrell, *Murphy-Durrell Diagnostic Reading Readiness Test*. World Book Co., Yonkers, N. Y., 1947.

<sup>22</sup> James I. Brown and G. R. Carsen, *Brown-Carsen Listening Comprehension Test*. World Book Co., Yonkers, N. Y., 1952.

subjects is to make a drive on the work type of reading ability. A recognition of this has caused makers of tests in reading to turn their attention in this direction in recent years. A number of excellent reading tests that provide useful analytical information concerning a number of work-study skills are available.

The *Gates Silent Reading Tests* are prepared in two series and two forms for use in the primary grades and in the intermediate grades. The primary tests are available in Types 1, 2, and 3, while four kinds, Types A, B, C, and D, of the intermediate test are available. Single exercises from each of Types 1, 2, and 3 are given as examples of the content of these tests.

The *Iowa Silent Reading Tests*, New Edition, Elementary, are among the more recent and comprehensive tests designed to provide a detailed and analytical measure of silent reading abilities. These new quick-scoring tests go beyond the general survey of two or three phases of silent reading ability. They cover a wide range of skills essential to effective reading of the work-study type. Naturally they do not succeed in measuring all of the major objectives of reading as outlined in the opening pages of this chapter.



These tests sample into numerous reading skills and into different subject-matter fields. For example, Test 1 consists of two articles, one dealing with science and the other with history content. Rate of reading is measured by these two samples under definite comprehension requirements. The pupil is told to read the articles as rapidly as possible and yet be able to answer certain comprehension questions based on the content. Test 2, *Directed Reading*, utilizes the same two articles of science and history content for an intensive check on the pupil's ability to comprehend certain questions and to locate their answers in the articles. Test 3, *Word Meaning*, contains two groups of exercises. The first group samples into general vocabulary, and the second into subject-matter vocabulary in mathematics, science, and social science arranged in cyclic order. Test 4 measures three phases of paragraph comprehension.



One of the more valuable reading skills measured by these tests is the ability to use the index for the purpose of locating information. This test, which is Part B of Test 6, is a good example of how this ability may be measured in a completely objective manner. It should also give to many teachers hints for the development of similar material for instructional purposes.









Excerpts from Gates Primary Reading Tests <sup>23</sup>

## TYPE 1. WORD RECOGNITION














	did	egg
	dog	two
	be	bed
	bag	she

	may	make
	come	milk
	horse	play
	hose	house

## TYPE 2. SENTENCE READING

<p>This is a cat. I</p> <p>This is a book. II</p> <p>This is a cup. III</p>	     
---	--

## TYPE 3. READING OF DIRECTIONS

   <p>1. Put an X on the ball.</p>	   <p>3. Draw a line under the little book.</p>
    <p>2. Put an X on the milk bottle.</p>	   <p>4. Draw a line from the pig to the tree.</p>

<sup>23</sup> Arthur I. Gates, *Gates Primary Reading Tests*. Published by Bureau of Publications, Teachers College, Columbia University, 1926 and 1931.

## Performance tests in silent reading

The *Betts Ready to Read Tests* include not only reading readiness tests but also tests for the diagnosis of difficulties for pupils who do not read normally. The tests for oculomotor and perception habits require the use of a series of slides and the *Betts-Keystone Telebinocular*,<sup>24</sup> a type of stereoscope that provides a scaled holder adjustable for various distances. The tests measure fusion, visual acuity, muscular balance, eye coordination, depth perception, and astigmatism. Their purpose is not to diagnose visual defects as a basis for prescription but to locate pupils who should be referred to eye specialists for examination and remediation.

The *Ophthalmograph*<sup>25</sup> is a binocular eye-movement camera used to obtain a simple and objective record of eye movements during the reading process. Information is provided on a film strip concerning the number of eye fixations, recognition span, regressive eye movements, rhythm, reading speed, and coordination of the eyes. Charts are provided for use in easy determination of total reading time for a given number of words. This procedure measures the eye mechanics of reading, and should ordinarily be supplemented by a test of reading comprehension.

The *Metronoscope*<sup>26</sup> is a device for exposing printed strips of reading matter at desired rates of speed and can be used either with individuals or small pupil groups for testing and drill purposes.

The *Durrell Analysis of Reading Difficulty* materials include a hand-operated tachistoscope, or device for exposing reading strips at desired rates, for use in determining word recognition and phrase comprehension. A test of oral reading measures phrase reading, voice, enunciation, expression, and general word skills, and is accompanied by questions to test comprehension.

## 6 CORRECTIVE EXERCISES IN READING

### Remedial drills for oral reading difficulties

Dearborn, Huey, Gray, Buswell, and many other early investigators, studying the problem of how to improve reading, noted that

<sup>24</sup> Distributed by Keystone View Co., Meadville, Penn.

<sup>25</sup> Distributed by the American Optical Co., Southbridge, Mass.

<sup>26</sup> *Ibid.*



there is a marked relationship between the rate and quality of children's reading and the control they have over their eye movements in reading. The meaning of eye movements may be readily understood by anyone who will take a position closely in front of and directly in the range of vision of a person engaged in reading. The observer will note that the reader's eyes do not move regularly and systematically forward as the reading progresses but that the movements are interspersed with pauses or fixation periods. It is during these pauses that the images of the words or groups of words are secured. Carefully conducted laboratory experiments reveal the fact that good readers make longer sweeps with the eyes, take in larger units of words, pause for much shorter periods, and rarely retrace material once covered. Gates <sup>27</sup> concluded that improper eye movements are probably the evidences of other types of reading disability that can be treated specifically. With the removal of the causes lying back of ineffective eye movements, the treatment of eye movements as such becomes unnecessary. On the whole this seems to be the most hopeful way of looking at the problem, since a great many teachers are qualified to administer types of remedial treatments that may be applied in the classroom but only a few have the technique or equipment for training the pupil in more effective eye movements. Gray himself recognized this practical aspect of a problem of remedial instruction in reading and suggested a number of excellent exercises designed to overcome specific difficulties in reading, many of which are indicated by the ineffectual eye movements of the pupil. One of these exercises is reproduced here to illustrate types of material adapted to remedying certain oral reading difficulties.

#### EXERCISE TO INCREASE ACCURACY OF RECOGNITION <sup>28</sup>

1. Words which a pupil failed to recognize accurately while reading were used in sentences at the end of each period, in order that he might associate them with their meaning. The words which repeatedly caused difficulty were then typewritten on cards and used in quick-perception drills, by presenting them as rapidly as they were recognized. Such words as *again, want, been, does, and heard* were fre-

<sup>27</sup> Arthur I. Gates, *The Improvement of Reading: A Program of Diagnostic and Remedial Methods*, Third edition. Macmillan Co., New York, 1947. p. 444.

<sup>28</sup> W. S. Gray, *Remedial Cases in Reading: Their Diagnosis and Treatment*. Supplementary Educational Monograph No. 22. University of Chicago Press, Chicago, 1922.

quently emphasized. As soon as a pupil was able to recognize a word readily, drill on it was discontinued. New words were added to the list as difficulties were encountered.

2. Words which a pupil confused because of their similarity in form were emphasized in drill exercises. These words included such groups as *thought, though, and through, there and where, then and when, now and how, and has, had, and have*. The words were used in sentences before they were presented in quick-perception drills. If unusual difficulties were encountered, words which were similar in form were presented together so that their differences could be studied.
3. Pupils who recognized isolated words accurately frequently made errors in recognizing the same words in phrases and sentences. In order to overcome this difficulty a word, such as *there*, was written on the board in several phrases or short sentences and the pupil was given opportunity to study them deliberately. As soon as he was able to recognize these phrases readily they were typewritten on cards and presented in quick-perception drills.

By practice in the use of such diagnostic reading devices as these and by training themselves in careful observation of their pupils, teachers can become adept in the detection of particular reading difficulties. Furthermore, they will soon find that with practice they can become proficient in the art of building drill exercises. The sample exercises selected from a great many suggested by Gray, Gates, and others will be found valuable guides in the preparation of such materials. By following the examples given here, the teacher can be practically certain that he is using reading drill material whose efficiency has been experimentally established.

### Remedial drills for work-study types of reading difficulties

Possibly one reason for the rather marked instructional emphasis on the work-study type of reading to the exclusion of reading of the leisure types is that it is reasonable to expect a considerable carry-over of skills from work-type reading to the other type, because of the large number of common skills involved. For example, skill in recognition of word meaning, which functions in work-study reading, is probably similarly effective when the individual is reading solely for pleasure.

A few illustrations of specific types of remedial exercises suited for use in silent reading of the work-study types are presented in the



following pages in the hope that they may serve as a guide to teachers interested in the development of material of this type. Only a few samples from each field can be furnished.

*Word recognition.* Exercises designed to develop skill in the recognition of new word meanings appear in a great many forms, as for example: (1) simple sentence completion; (2) agent-action; (3) action-agent; (4) action-effect; (5) effect-action; (6) identification; (7) opposites; (8) similars; (9) description; and (10) phrasing.

*Location of information in books.* A significant factor in the child's use of reading for work-study periods is his ability to locate information in books. The following suggestions may prove helpful:

1. To develop in pupils an ability to use the index, children should:  
(a) be taught the alphabet; (b) be drilled in arranging words in alphabetical order; (c) be drilled in finding answers to questions by use of the index; (d) be asked to prepare indexes for books not provided with them.
2. To develop the ability to use a table of contents, pupils should: (a) be assigned lessons by topic or titles; (b) find the assigned lessons in the text by means of the table of contents; (c) find additional sources of information on the assignment in the library.

*Organization of material.* The ability to organize what is read is a necessary part of the equipment of everyone who expects to become a good student. Organization of reading materials calls for a superior type of judgment. The following suggestions will aid teachers in developing a variety of types of practice in organization:

1. Practice in deciding upon the main thought in the paragraph or topic.
2. Drill in outlining a study, an assignment, a reference reading, a poem.
3. Practice in analyzing the organization of selections.
4. Practice in restating the substance of a difficult passage to convey the same idea in simplified form.
5. Practice in selecting the most appropriate title for a selection.

## Other remedial materials in reading

The recent analytical work of Betts, Durrell, Gates, Gray, and others opens up great possibilities for diagnostic and corrective work. No longer need diagnosis in reading be confined to such vague and general qualities as rate and comprehension of word meanings. In

fact, it now becomes quite apparent that many of the so-called diagnostic tests in this field are not at all suited for the specific types of diagnosis required in the identification of reading disabilities. While considerable progress has been made in the last decade in the more exact analysis and identification of underlying causes of reading disability, the development of adequate initial instructional materials and corrective devices has not kept pace with the analytical work. The next decade is almost certain to see much progress along these lines.

Commercial materials designed for instructional testing purposes and remedial uses in reading are available in many different forms. At present there are almost countless drill books and workbooks having for their purpose the development of silent reading skills of the work-study type. However, this material has mainly emphasized the problems of teaching beginners to read by some particular method rather than that of providing corrective treatment for some basic disability.

### Topics for Discussion

1. State your idea of the relative importance of the receptive and the expressive language arts skills? Would human society be able to dispense with either?
2. Evaluate the school importance of listening as a receptive language skill area. Is listening as important socially as silent reading abilities or the study skills? Defend your position.
3. Why would you expect reading disability to be reflected in classroom achievement?
4. In what specific ways does modern life place a particular burden on the ability to read rapidly and well?
5. Prepare a comprehensive list of oral reading objectives.
6. Check the list of Essential Skills Involved in Work-Study Types of Reading against the skills specified for measurement in the Iowa Silent Reading Tests.
7. Summarize your position with respect to the relative placement of instructional emphasis on oral and silent reading.
8. Prepare a set of suggestive drill exercises for the development of skill in the use of the index.
9. Classify the major reading defects according to type.
10. Are improper eye movements in reading causes of reading deficiencies or merely evidences of such defects?



11. Prepare suggestions for the increase in the pupil's span of attention in reading.
12. Is there any special reason to assume that remedial practices that are helpful in curing oral reading defects may not also be useful in correcting silent reading deficiencies?

## Selected References

- ANDERSON, IRVING H., AND DEARBORN, WALTER F. *The Psychology of Teaching Reading*. New York: Ronald Press Co., 1952.
- BETTS, EMMETT A. *Foundations of Reading Instruction with Emphasis on Differential Guidance*. New York: American Book Co., 1946.
- BETTS, EMMETT A. *The Prevention and Correction of Reading Difficulties*. Evanston, Ill.: Row, Peterson and Co., 1936.
- BOND, GUY L., AND BOND, EVA. *Teaching the Child To Read*. New York: Macmillan Co., 1943.
- BROOM, M. E. *Educational Measurements in the Elementary School*. New York: McGraw-Hill Book Co., Inc., 1939. p. 182-204.
- BROWN, JAMES I. "The Construction of a Diagnostic Test of Listening Comprehension." *Journal of Experimental Education*, 18:139-46; December 1949.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 317-23, 333-35, 567-620.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p. 140-43, 336-79.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938. p. 79-80, 124-40, 155-56.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. p. 247-57, 501-71.
- DOLCH, EDWARD W. "Testing Reading with a Book." *Elementary English*, 28:124-25; March 1951.
- DURRELL, DONALD D. *Improvement of Basic Reading Abilities*. Yonkers, N. Y.: World Book Co., 1940.
- FINCH, F. H., AND GILLENWATER, V. W. "Reading Achievement Then and Now." *Elementary School Journal*, 49:446-54; April 1949.
- GATES, ARTHUR I. *The Improvement of Reading: A Program of Diagnostic and Remedial Methods*. Third edition. New York: Macmillan Co., 1947. Chapter 3.

- GATES, ARTHUR I. "The Measurement and Evaluation of Achievement in Reading." *The Teaching of Reading: A Second Report*. Thirty-Sixth Yearbook of the National Society for the Study of Education, Part I. Bloomington, Ill.: Public School Publishing Co., 1937. Chapter 12.
- GATES, ARTHUR I., chairman. *Reading in the Elementary School*. Forty-Eighth Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1949.
- GATES, ARTHUR I., AND OTHERS. *Methods of Determining Reading Readiness*. New York: Bureau of Publications, Teachers College, Columbia University, 1939.
- GERBERICH, J. RAYMOND. "The First of the Three R's." *Phi Delta Kappan*, 33:345-49; March 1952.
- GRAY, WILLIAM S. "Contributions of Research to Special Methods: Reading." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1938. Chapter 7.
- GRAY, WILLIAM S. "Reading." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 965-1005.
- GRAY, WILLIAM S., chairman. *The Teaching of Reading: A Second Report*. Thirty-Sixth Yearbook of the National Society for the Study of Education, Part I. Bloomington, Ill.: Public School Publishing Co., 1937.
- GREENE, HARRY A., AND GRAY, WILLIAM S. "The Measurement of Understanding in the Language Arts." *The Measurement of Understanding*. Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1946. p. 189-200.
- HARRIS, ALBERT J. *How To Increase Reading Ability: A Guide to Individualized and Remedial Methods*. Second edition. New York: Longmans, Green and Co., 1947.
- HARRISON, M. LUCILE. *Reading Readiness*. Boston: Houghton Mifflin Co., 1936.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. p. 95-117, 152-56, 167-81.
- LESTER, JOHN A., AND LINDQUIST, E. F. "Examinations in English." *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Co., 1936. p. 381-410.
- McKEE, PAUL. *The Teaching of Reading in the Elementary School*. Boston: Houghton Mifflin Co., 1948.
- MONROE, MARION. "Diagnosis and Treatment of Reading Disabilities." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1935. Chapter 12.



- MONROE, MARION, AND OTHERS. "Diagnostic and Remedial Procedures in Reading." *Educational Record*, 19:105-13, Supplement No. 11; January 1938.
- MONROE, MARION, AND OTHERS. *Remedial Reading*. Boston: Houghton Mifflin Co., 1937.
- MORSE, HORACE T., AND McCUNE, GEORGE H. *Selected Items for the Testing of Study Skills*. National Council for the Social Studies, Bulletin No. 15. Washington, D. C.: National Education Association, September 1940.
- RUSSELL, DAVID H. "Evaluation of Pupil Growth in and through Reading." *Reading in the Elementary School*. Forty-Eighth Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1949. Chapter 14.
- SMITH, DORA V. "Recent Procedures in the Evaluation of Programs in English." *Journal of Educational Research*, 38:262-75; December 1944.
- SPACHE, GEORGE. "A Comparison of Certain Oral Reading Tests." *Journal of Educational Research*, 43:441-52; February 1950.
- SPACHE, GEORGE. "The Construction and Validation of a Work-Type Auditory Comprehension Reading Test." *Educational and Psychological Measurement*, 10:249-53; Summer 1950.
- THORNDIKE, EDWARD L. *A Teacher's Word Book of the Twenty Thousand Words Found Most Frequently and Widely in General Reading for Children and Young People*. New York: Bureau of Publications, Teachers College, Columbia University, 1931.
- TRAXLER, ARTHUR E. "Measurement in the Field of Reading." *English Journal*, 38:143-49; March 1949.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapters 8-9.
- WITTY, PAUL. "Approach to Better Reading: An Evaluation." *Educational Administration and Supervision*, 25:81-92; February 1939.
- WITTY, PAUL A., AND KOPEL, DAVID. "Evaluating Reading and Remedial Reading." *English Journal*, 26:449-58; June 1937.
- WITTY, PAUL A., AND KOPEL, DAVID. "Preventing Reading Disability: The Reading Readiness Factor." *Educational Administration and Supervision*, 22:401-18; September 1936.
- WOOD, BEN D., AND HAEFNER, RALPH. *Measuring and Guiding Individual Growth*. New York: Silver Burdett Co., 1948. p. 285-91.
- WRIGHTSTONE, J. WAYNE. "Diagnosing Reading Skills and Abilities in the Elementary School." *Educational Method*, 16:248-54; February 1937.

## ***Measuring and Evaluating in the Expressive Language Arts***

THIS CHAPTER presents a summary of the following points concerning measurement at the elementary-school level in the fields of expressive language arts—language, spelling, and handwriting:

- A. Analysis of specific language skills.
- B. Measurement of oral and written language skills.
- C. Remedial instruction in language.
- D. Construction of spelling tests.
- E. Diagnosis and remediation in spelling.
- F. Measuring quality and rate of handwriting.
- G. Diagnosis and remediation in handwriting.

The expressive language arts, consisting of language, spelling, and handwriting, are discussed in this chapter. These, together with reading and the work-study skills discussed in the preceding chapter, round out the language arts subjects ordinarily taught in the elementary school.

### **1 IDENTIFICATION OF LANGUAGE ABILITIES**

#### **The social importance of language**

The importance of communication in everyday life naturally gives the language arts a place of unusual prominence among the instructional problems faced by the classroom teacher. It is significant that



language, which was one of the first subjects to be measured, is one of the slowest to respond to analysis, diagnosis, and remedial treatment. Possibly this is due in part to the formal methods of instruction in this subject followed by many teachers. It is more likely to be due, however, to the sheer complexity of the subject itself, and the many forms in which it expresses itself.

In this discussion, language skill is considered to mean facility in the use of the proper language habits and forms essential to effective intercommunication at a particular cultural level. Such a point of view makes reasonably clear the problems of the language teacher. Language skills arise, as do other specific skills, through the proper exercise of the desired habits. It goes without saying that proper exercise is possible only when proper identification of the habits has taken place. One cannot indulge in exercises until he knows what to exercise. Hence the habits upon which language skill depends must be identified, and much carefully constructed instructional and drill material must be provided. This in itself will serve two useful purposes. First, the use of good language drill material insures that the pupil will have experience in making the correct response to selected language situations either with or without the assistance of such formal grammar instruction as may be applied. Second, the use of such material sets up in the pupil's mind an attitude toward language error. A definite consciousness toward language errors must be developed. When a person becomes sensitive to these, and an expression such as "he don't" causes a reaction similar to that created in a cat confronted by a strange dog, he is on the way to rapid improvement in his language habits.

### Analysis of language skills

It must be evident from what has been presented in earlier chapters in this book that an accurate analysis of the underlying skills in language is necessary before any significant program of diagnostic and corrective work can be undertaken. In the past, certain general language abilities have been identified for measurement purposes, such as language usage, grammar, and composition. In more recent years, however, there has been an effort to reduce language in general to its more elementary or basic skills.

The task of clarifying the statements and purposes of language

(English) instruction and of analyzing and identifying the basic language skills is not one that will be successfully accomplished by any one person. At the outset it must be recognized that there are many conditions under which language functions. There is undoubtedly a language of impression, or comprehension, as well as a language of expression. The former is the aspect of language that is given particular attention in reading instruction. The latter, the language of expression, is the phase usually meant by the term "language ability," and is the phase that receives special attention in language instruction.

The program of instruction in language expression must equip the child to engage successfully in certain speaking and writing activities wherever he may encounter them, either in school or out of school. In a discussion of the program in language arts, McKee<sup>1</sup> pointed out that the school is responsible for providing definite instruction and experiences in each of the following important speaking and writing situations encountered in school and in life outside the school: (1) taking part in conversation and discussions, (2) using the telephone, (3) taking part in meetings, (4) giving reports, (5) telling and writing stories, (6) giving reviews and reports, (7) giving directions and explanations, (8) making announcements, (9) giving descriptions, and (10) writing letters. Participation in these language activities means that the child must make use of a great many specific abilities, understandings, skills, and attitudes, which, from the language instructional point of view, represent the objectives of the language curriculum.

The accompanying outline of language outcomes and objectives is a compilation and an adaptation of material from several sources. To develop a complete and perfect outline of language objectives is almost certainly a hopeless task. In spite of certain logical and psychological shortcomings which this outline may possess, it nevertheless gives the teacher helpful suggestions for the identification of useful language skills. The teacher and the student will, of course, wish to revise such an outline from time to time to keep it in line with the best research evidence in the field.

<sup>1</sup> Paul McKee, "An Adequate Program in the Language Arts." *Teaching Language in the Elementary School*. Forty-Third Yearbook of the National Society for the Study of Education, Part II. Department of Education, University of Chicago, Chicago, 1944. Chapter 2.



LANGUAGE OUTCOMES AND OBJECTIVES<sup>2</sup>

## I. Oral Language

## A. General Outcomes of Oral Language

1. To form correct habits of articulation, enunciation
2. To assume proper and pleasing body position and mannerisms when speaking
3. To use common courtesy in social groups
4. To speak with feeling, reflecting meaning, thought, and interest
5. To learn how to locate and give information
6. To think while speaking
7. To develop sentence sense
8. To pronounce correctly such words as are used
9. To learn how to acquire new words
10. To speak at a rate suitable to conditions
11. To use voice of suitable clarity and loudness
12. To listen courteously and critically
13. To recognize listening as an essential part of every conversational situation
14. To develop judgment of what is suitable to talk about in conversational groups

## B. Special Oral Language Outcomes and Situations

1. To relate anecdotes and incidents interestingly
2. To make necessary simple announcements
3. To participate with ease in conversation
4. To take an active part in arguments and debates
5. To learn to disagree or argue courteously
6. To learn to listen, summarize, and report activities, events, news items, instructions

<sup>2</sup> Adapted from the following sources: (1) Maude McBroom, *The Course of Study in Written Composition for the Elementary Grades*. University of Iowa Monographs in Education, First Series, No. 10. University of Iowa, Iowa City, 1928. (2) *Iowa Elementary Teachers' Handbook: Oral and Written Language*. State Department of Public Instruction, Des Moines, 1944. (3) Harry A. Greene and H. L. Ballenger, *Manual of Instructions: Iowa Language Abilities Tests*. World Book Co., Yonkers, N. Y., 1948. (4) Paul McKee, *Language in the Elementary School*. Houghton Mifflin Co., Boston, 1939. (5) M. R. Trabue, chairman, *Teaching Language in the Elementary School*. Forty-Third Yearbook of the National Society for the Study of Education, Part II. Department of Education, University of Chicago, Chicago, 1944. (6) Harry A. Greene and William S. Gray, "The Measurement of Understanding in the Language Arts." *The Measurement of Understanding*. Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. University of Chicago Press, Chicago, 1946. p. 176-89.

7. To learn to react properly to social responsibilities, as an introduction, meeting a stranger
8. To develop ability to assume an active part in school activities, as committee meetings, associations, classroom dramatizations, plays
9. To learn to take the proper auditor-speaker attitudes
10. To use the telephone properly

## II. Written Language

### A. General Outcomes, Knowledges, and Skills Peculiar to Written Composition

1. To answer letters promptly
2. To develop judgment of the suitability of content for special situations, such as friendly letters, business letters, letters of sympathy
3. To learn to use correct form in writing business and social letters, notes, invitations
4. To learn to fill in common forms, blanks
5. To acquire skill in writing notices, announcements, and advertisements, telegrams
6. To show interest and skill in doing creative writing, such as stories, plays, editorials, diaries
7. To acquire skill in making outlines from content material
8. To record minutes of meetings, dictations by teacher
9. To acquire skill in evaluating, organizing class and lecture notes
10. To prepare an accurate and comprehensive bibliography

### B. Knowledges and Outcomes Peculiar to All Written Work

1. To write legibly and rapidly
2. To spell correctly socially useful forms
3. To utilize proper manuscript forms
4. To use proper outline forms
5. To punctuate written work correctly
6. To capitalize correctly

## III. Knowledges and Skills Common to Both Oral and Written Language

### A. General Outcomes

1. To develop a sincere desire to speak and to write correctly
2. To develop a sensitivity to error in speaking and writing
3. To learn to use sources of information, as dictionary, encyclopedias, reference books



4. To develop skill in producing variety in sentence structure
5. To identify types of sentences so that voice and punctuation can clearly indicate meaning
6. To learn to expand the meaning vocabulary

B. Correct Usages

1. To master the most important grammatical usages, as pronouns, verb forms, subject-predicate relationships, redundancy, double negatives, antecedents

C. Rhetorical Skills

1. To develop sentence sense
2. To develop variety in sentence structure
3. To avoid faults in sentence structure, as useless introductory words, phrases, loose use of connectives
4. To organize ideas in sentences so that the sentence says exactly what is meant
5. To organize ideas and sentences in a paragraph around a single topic
6. To organize ideas in a paragraph in proper sequence
7. To avoid use of overworked words
8. To develop through use a vocabulary rich in color, accuracy, conciseness, suitability, variety
9. To stimulate interest through use of contrasts, concreteness, variety, simile-metaphors

## Oral language skills

The foregoing catalogue of outcomes indicates that language as a means of verbal expression appears in two main forms, oral and written. Success in the use of oral language depends in the first place on the ability of the speaker to so choose, arrange, and enunciate his words as to affect his hearers as he intends. In order to guarantee success in the operation of these skills, the pupil must be given training and practice in thinking and talking under audience conditions. In this training, emphasis must be placed on the development of a pleasant speaking voice, a gracious attitude, a clear enunciation of words, an avoidance of common language errors, care in the selection of words, a careful selection and organization of ideas, and skill in the clothing of his thoughts in the proper words so that he may affect his hearers as he intends by leading their thinking along prescribed channels.

It is equally imperative that attention be given to the development of proper skills and attitudes on the part of the listener. Effective audience-speaker reactions are the result of an interplay of factors arising from the fact that (1) something of value is being communicated, (2) between persons appreciative of the values, but (3) possessing them at different levels of control. Thus the essential elements of the audience situation are present. The speaker has a message. The audience is ready to listen courteously and critically because useful and interesting information is being communicated.

### Written language skills

The problem of written language takes a threefold form, although this is not apparent from the outline of outcomes. The first involves the formal or mechanical factors, such as writing, spelling, punctuation, form, and general appearance. The second treats of certain grammatical factors, such as common errors in language form and sentence structure and form. The third is concerned with the more subtle elements of composition, the rhetorical factors involving the questions of choice of words, quality of interest innate in the material, and logical organization of the subject matter both within the sentence and the larger units. In the first two phases of the problem of written language, the factors are more generally uniform in their manner of affecting readers. However, there is greater difficulty in predicting the effect of the third phase on the reader. These mechanical and grammatical elements constitute in a way the raw material of written expression. The rhetorical factors are the results of the manner in which these raw materials are put together. They are the factors that make for appeal, originality, style, and distinctiveness in written expression. The mechanical and grammatical factors are relatively tangible, objective, and measurable. The rhetorical factors are more intangible, more difficult to identify and to measure, and thus far some of these elements have eluded the best efforts to measure them objectively.

## 2 MEASUREMENT AND DIAGNOSIS OF LANGUAGE ABILITIES

### Oral language scales

An examination of the foregoing outline of language outcomes makes it clear that oral language ability is made up of many related



general and specific abilities. It is also equally obvious that from the standpoint of its social utility oral language is extremely important. Yet measurement of oral language abilities is strangely limited. In fact, so far as the writers know, there are no adequate standardized instruments for the measurement of oral composition that will stand inspection in the light of present-day criteria. Some progress has been made in the evaluation of techniques for measuring the improvement in oral composition, but thus far no practical way of making the results available to the classroom teacher has been devised.

## Diagnosis of oral language disabilities

Considerable progress in the identification of oral language disabilities and in the development of remedial procedures in oral expression has been made by investigators in the field of speech. It is clear, however, that any classification of speech disorders must necessarily be conditioned by individual points of view. For example, Blanton<sup>3</sup> recognized four fundamental speech disorders: (1) delayed speech, (2) oral inactivities, (3) letter substitutions, and (4) stuttering. On the other hand, Mulgrave treated the problems of speech pathology under the three main headings indicated in the accompanying outline, which is adapted from three chapters in her textbook, *Speech for the Classroom Teacher*.<sup>4</sup>

### PROBLEMS OF SPEECH PATHOLOGY

#### I. Functional Speech Disorders

- A. Baby talk (infantile speech sounds and substitutions)
- B. Defective phonation (faulty production of speech sounds)
  - 1. Inorganic lisping (impure production of sibilant sounds)
  - 2. Lingual protrusion (misplacement of tongue)
  - 3. Lateral emission (due to formation of teeth and tongue placement)
  - 4. Nasal emission (poor control of soft palate causing sound to be emitted through the nose)

<sup>3</sup> Smiley Blanton, "Problems and Methods in the Correction of Defective Speech." *Speech Training and Public Speaking for Secondary Schools*. Report of special committee of the National Association of Teachers of Speech. Century Co., New York, 1925.

<sup>4</sup> Dorothy I. Mulgrave, *Speech for the Classroom Teacher*, Revised edition, Prentice-Hall, Inc., New York, 1946.

## C. Vulgar speech

1. Foreign accent (sound omissions, substitutions, intonations due to influence of a foreign language)
2. Regional dialects (conspicuous speech deviation that labels speaker geographically)

## II. Organic Speech Disorders

- A. Organic lisping (due to malformation of jaws or to failure of jaws to meet properly)
- B. Tongue-tie (movement of the tongue impeded)
- C. Cleft palate (defective palate or roof of mouth)
- D. Chronic hoarseness of voice (may be due to pathological condition, misuse, or to a neurotic condition)
- E. Nasality (caused by too large a proportion of nasal resonance)
- F. Denasality (too little nasal resonance resulting from chronic catarrh, sinus infection, adenoids)

## III. Emotional Disorders

- A. Stammering (any habitual hesitation or repetition in forming speech sounds)
- B. Neurotic lisping (persists because individual desires to keep it in spite of lack of physical cause)
- C. Neurotic hoarse voice (may be due to nervousness or hysteria)

Travis<sup>5</sup> preferred to group speech disorders under the three following heads: (1) disorders of rhythm in verbal expression, (2) disorders of articulation and vocalization, and (3) disorders of symbolic formulation and expression.

*Disorders of rhythm in verbal expression.* This group of speech disorders includes stammering and stuttering, which Travis considered basically similar. While the number of serious cases of stuttering is not actually very great, the effect on the individual is so serious that it is important for teachers to have some idea of the nature and extent of this disorder. Careful surveys indicate that approximately one pupil per hundred will be a stutterer, with the boys far outnumbering the girls in this speech handicap. Apparently there is no very definite relationship between stuttering and the mental ability of the pupil.

Since the classroom teacher, no matter how great may be his interest in the nature and the causes of stuttering, can do very little about

<sup>5</sup> Lee E. Travis, *Speech Pathology*. D. Appleton-Century Co., Inc., New York, 1931. p. 37.



it, the important thing in connection with instruction in oral language is for him to develop the proper understanding of and sympathy for the stutterer's outlook on life.

*Disorders of articulation and phonation.* Normal speech implies the existence of adequate speech equipment in the physical sense capable of responding to the proper stimuli. The production of speech sounds calls for the most accurate coordination of the physical and mental aspects of the speech mechanism.

Under this category of disorders of articulation and phonation are classified all of the defects found in enunciation and voice production, including delayed development of speech. Travis<sup>6</sup> pointed out that in this field there are two types of speech defects, (1) functional defects that are due to bad training, and (2) organic defects that come from injuries or from faulty or abnormal development of the brain or other organs related to speech. Many of this particular class of speech defects arise from such organic difficulties as abnormal development of the tongue, cleft palate and harelip, abnormal development of the jaws and teeth, adenoids, and defective hearing.

The treatment for most of these disorders of articulation and phonation involves medical, mental, hearing, and speech examinations. Since these are generally highly technical in character, they should probably be undertaken only by the trained specialist in each field.

*Disorders of symbolic formulation and expression.* Travis<sup>7</sup> defined disorders of symbolic formulation and expression as consisting essentially of "a lack of power to execute with ease acts connected with articulated speech and the comprehension of spoken words." The location of these defects is largely a clinical rather than a classroom problem. Accordingly, the teacher, upon the discovery of any cases among his pupils who are unable to articulate or to comprehend the spoken word, should immediately refer them to a clinical expert.

## Skills peculiar to written language

The catalogue of language outcomes presented on pages 422 to 424 is a reasonably satisfactory classification for the purpose of contrasting two major types of verbal expression, but it seems inadequate when considered from the point of view of the complete

<sup>6</sup> *Ibid.* p. 37-38, 196, 211.

<sup>7</sup> *Ibid.* p. 232.

identification and analysis of the specific underlying skills upon which verbal expression depends. For this more exacting purpose a classification based on such units of language form as the word, the sentence, the paragraph, and the composition unit, and on certain general mechanical factors, is superior. In order to present a more concrete idea of the types of abilities called into play at each of these levels of language skill, the accompanying detailed outline is given.

#### DIAGNOSTIC OUTLINE OF LANGUAGE SKILLS

- A. Words—Skill in the spelling, choice, use, and definition
  - 1. Spelling—ability to spell certain socially useful words
    - a. Contractions
    - b. Abbreviations
  - 2. Choice of words
    - a. Same
    - b. Opposite
    - c. Exact word for meaning
    - d. Variety
    - e. Meaningful words
    - f. Minimum number of words
    - g. Semantic variations in meanings
  - 3. Correct usage
    - a. Verbs
    - b. Pronouns
    - c. Modifiers
    - d. Nouns
  - 4. Use of dictionary
    - a. Alphabetizing
    - b. Use of guide words
    - c. Selection of meaning
    - d. Using pronunciation keys
- B. Sentences—Skill in the use, form, structure, and organization
  - 1. Form
    - a. Complete, coherent, unified
    - b. Variation in beginning
    - c. Variation in length
  - 2. Kind
    - a. Declarative
    - b. Interrogative
    - c. Exclamatory



3. Structure
  - a. Simple, compound, complex
  - b. Subject and predicate
  - c. Variety in structure
  - d. Language usage—avoidance of slang and foreign expressions, faulty expressions, double negatives
4. Organization
  - a. Logical sequence of ideas
  - b. Variety for interest
- C. Paragraphs—Skill in the form, structure, and organization
  1. Form
    - a. Indentation
    - b. Initial and terminal line length
    - c. Length
  2. Structure
    - a. Unity
    - b. Coherence
  3. Organization
    - a. Outline
    - b. Logical sequence of ideas
- D. Letter writing—Skill in selection of content and in use of form and mechanics in
  1. Business letters
  2. Social letters
  3. Informal notes
  4. Formal notes
- E. Outline form
  1. Organization
  2. Capitalization
  3. Punctuation
- F. Bibliographical form
  1. Arrangement for unpublished material
  2. Arrangement for published material
  3. Capitalization
  4. Punctuation
- G. General mechanical factors—Skill in control of
  1. Capitalization
    - a. Initial words in sentences
    - b. Proper nouns

- c.* Proper adjectives
  - d.* Titles of honor and respect
  - e.* Important words in titles of stories, articles, etc.
2. Punctuation
  - a.* End
  - b.* Series
  - c.* Quoted matter
  - d.* Special situations
3. Margins
  - a.* Top, bottom, sides
  - b.* Indentation
4. Handwriting
  - a.* Legibility
  - b.* Speed
5. Abbreviations
  - a.* Titles
  - b.* Other situations
6. Hyphenations
  - a.* Compound words
  - b.* Ends of lines

In spite of the detail of this outline and the number of specific skills that contribute directly to language ability, the reader will immediately recognize certain significant weaknesses. Many of the skills are identified only in a very general way. The recognition of choice of words as a significant language skill is approximately equal to stating that addition is an important skill in arithmetic. Much more definite information is necessary before all of the details of a constructive program of language improvement can be developed. Just as it is necessary to identify the socially useful situations and facts, or the most useful words in spelling, it is necessary to identify the skills that have the greatest social usefulness in language situations. Much excellent work has been done on the problem of determining a minimal spelling (writing) vocabulary based on social utility. Similar work must be done from the standpoint of language situations. Until this is accomplished, workers interested in the development of diagnostic exercises in oral and written language must turn to other sources for valid test and drill materials.



## Measures of general merit of written composition

The measurement of general merit of written composition, while dating well back into the history of educational measurement, has not responded to efforts to improve it in proportion to the attention it has received. This difficulty comes from the great complexity of the skills involved in producing merit in written language, and from the vagueness with which these skills have been recognized. Historically, the *Hillegas Composition Scale* antedates most other attempts to measure educational products. Not only has this scale accomplished much good through the stimulation of interest in the more accurate measurement of written composition, but it is still a usable instrument in its present form, the *Thorndike Extension to the Hillegas Scale for the Measurement of Quality in English Composition by Young People*.

Among the more useful of the currently available scales for the measurement of composition quality is the *Willing Scale for Measuring Written Composition*. This scale is made up of eight specimens of composition all written on the topic, "An Exciting Experience." Through the definite recognition of the relation of form errors to the general quality of written work this scale increases its usefulness. Its value is also enhanced through the very clear directions for the collection of compositions for survey purposes. An excellent list of interesting topics is also suggested as the basis for the written work. The use of such standardized lists of topics and the control of conditions under which the writing takes place add distinctly to the reliability with which written composition abilities may be measured.

## Standard tests in grammar and usage

While the measurement of the common grammatical usages is not confined to the field of written language, the very nature of the subject matter itself makes it necessary to measure it in written form. For those who believe that there is a formal as well as a functional aspect of usage, the *Kirby Grammar Test* still meets the need for this type of test in the seventh and eighth grades. This test measures the ability of the pupil to select the correct one of two usages in a sentence situation and to recognize the correct grammatical reason for his choice. The content of the usage exercises is based on a rather old study of the typical errors of children. Numerous comparisons of scores on usage and grammatical principles right

on this test fail to show a significant positive relationship. Somewhat in contrast with this test, the *Iowa Grammar Information Test*,<sup>8</sup> usable also in the seventh and eighth grades, measures purely informational aspects of English grammar in eighty specific situations.

### Analytical measurement of language abilities

In addition to the foregoing tests, each of which presents only limited analytical possibilities in the measurement of language, there are three or four others that should be mentioned. In the light of the criteria for diagnostic measurement that have been set up in this volume, most of these tests fall short of being really diagnostic. In fact, it is very doubtful if there are any truly diagnostic tests in the language field. The *Franseen Diagnostic Tests in Language* are diagnostic only to the extent that they identify difficulties dealing with pronouns, verbs, and varied constructions. In spite of this fact they are very useful tests for survey purposes in Grades 3 to 8. The Language Section of the *Stanford Achievement Tests* deals with usage only. The *Unit Scales of Attainment in Language* are considerably more comprehensive, dealing with three aspects of language ability: capitalization, punctuation, and usage.

The *Iowa Language Abilities Tests* represent a program of analytical measurement in language that offers more than ordinary breadth in grade and content coverage. The Primary test is designed and standardized for use in Grades 1, 2, and 3. The following types of abilities are measured: filling in forms, conversation, oral composition, telephone conversation, correct usage, recorded composition, miscellaneous social usages, and letter writing. Only parts of filling in forms are attempted in the first grade. Letter writing is tested by recognition in Grades 2 and 3 only. The tests may be administered as group measures in the second semester of the first grade and in the second and third grades. In general, the pupils' reactions are simple and objective. A feature of the test is the use of a single small test booklet for both forms of the test. The column of the directions followed in the Examiner's Manual determines the form of the test that is administered.<sup>9</sup>

The Elementary test and the Intermediate test follow the general

<sup>8</sup> Fred D. Cram and H. A. Greene, *Iowa Grammar Information Test*. Bureau of Educational Research and Service, State University of Iowa, Iowa City, 1935.

<sup>9</sup> H. A. Greene and Lou A. Shepard, *Iowa Primary Language Test*. Bureau of Educational Research and Service, State University of Iowa, Iowa City, 1936.



pattern of the content of the original *Iowa Elementary Language Tests*, which these instruments displace. The Elementary battery is for use in Grades 4 to 6; the Intermediate test is for use in Grades 7 to 9. Illustrations of the techniques of testing word meaning, capitalization, and punctuation are given below.

Test C of the *Iowa Basic-Skills Tests* is a language test of considerable analytical power. The Elementary test is designed for use in Grades 3, 4, and 5 and the Advanced test for Grades 5 to 9.

### Excerpts from Iowa Language Abilities Test <sup>10</sup>

#### TEST 2. WORD MEANING

**DIRECTIONS:** One of the five numbered words in each exercise has almost the same meaning as the first word. One of the words means almost exactly the opposite of the first word. Find the word which means the *same* as the first word. Note its number. Then fill the answer space under **SAME** at the right which has the same number as this word. Next find the number of the word which means the *opposite* of the test word and fill the answer space under **OPPOSITE** at the right which is numbered the same as this word. Study the samples below.

##### SAMPLES:

<b>A</b>	<b>High</b>	1 slim	2 tall	3 short	4 dark	5 large	.....	s	1	2	3	4	5	o	1	2	3	4	5
<b>B</b>	<b>Cold</b>	1 sick	2 warm	3 tired	4 chilly	5 lonely	.....	s	1	2	3	4	5	o	1	2	3	4	5

#### TEST 6. CAPITALIZATION

**DIRECTIONS:** In some of the following sentences a word is written with a small letter which should begin with a capital letter, or a word is written with a capital which should not be capitalized. Each such word is numbered. Notice the number of this word. Then fill in the answer space at the right which is numbered the same as the word in the sentence which is not written correctly. Some of the sentences are written correctly. If a sentence is correct, fill in the answer space under **N**. The samples are answered correctly. Do the remaining exercises in a similar manner.

<b>SAMPLES:</b>	<b>A</b>	1 my	2 mother	3 came	4 home	early.	.....	A	1	2	3	4	N
	<b>B</b>	1 Did	2 Jim	3 take	4 the	car?	.....	B	1	2	3	4	N
	<b>C</b>	1 I	2 live	3 in	4 the	Country.	.....	C	1	2	3	4	N

#### TEST 7. PUNCTUATION

**DIRECTIONS:** In each of the following sentences certain words are printed in type like *this*. This means that you are to look for some punctuation mark which may be needed before, within, or after this word. In some cases no punctuation is needed. Study each sentence and decide which punctuation mark, if any, is needed. In the answer spaces at the right, fill in the space under the correct punctuation mark to use in the sentence at the place indicated by the word. If you think that no punctuation is needed, fill the answer space under **N**. The samples are answered correctly.

SAMPLES:	A	We saw an apple on the tree	.....	A	1	2	3	4	N
	B	Are you coming with me?	.....	B	1	2	3	4	N

## 3 REMEDIAL INSTRUCTION IN LANGUAGE

Remedial instruction in language will be effective only to the extent that pupils are made aware of the social importance of correct

<sup>10</sup> Harry A. Greene and H. L. Ballenger, *Iowa Language Abilities Tests*: (1) Test 2, Elementary; (2) Test 6 and Test 7, Intermediate. Published by World Book Co., 1948.

language usage and are led to develop a desire to make use of the best forms of expression and to formulate correct habits of usage. Language tests of the analytic types should aid in the developing of a self-critical attitude on the part of the pupil which naturally leads to the desire to acquire correct habits of expression.

## Remedial suggestions on punctuation and usage

Specimen types of remedial exercises in language are not presented here for two reasons. In the first place, there are countless excellent practice and drill books in the language field that provide adequate experience in the important skill areas. In the second place, the parallel between the desirable types of language drills and the types of exercises used in the tests to reveal the presence or absence of the skills is very close.

A helpful organization of diagnostic and remedial suggestions is presented in the *Manual for Interpreting the Iowa Language Abilities Tests*. Reproductions of the suggestions for the improvement of punctuation and language usage are presented on pages 436 and 437 as examples of this type of material. Similar suggestions for the improvement of work in spelling, word meaning, sentence sense, and capitalization are also given in this manual.

## Remedial exercises on form and appearance

The social importance of form and appearance in letter writing, one of the most frequent social uses of written composition, places a premium on such skills in this field. Exercises of the following types are suggested as useful drills on letter form.

### LETTER FORM EXERCISE

Arrange, capitalize, and punctuate properly the following items that go to make up the heading and salutation of a letter:

gentlemen  
236 erie avenue  
columbus city mo  
july 18 1926  
the sanford and morris company

(Use your own home city and address in the heading of the letter.)



TABLE 31. Diagnostic and remedial chart: language usage and punctuation<sup>11</sup>

LANGUAGE USAGE

POSSIBLE CAUSES OF LOW TEST SCORES	ADDITIONAL EVIDENCE OF DEFICIENCY	SUGGESTED REMEDIAL TREATMENT
1. Failure to comprehend the testing technique.	1. Lack of understanding of method of recording responses to items.	1. Prepare and use drill exercises similar to those used in the test. Work with individual pupil until he understands the technique.
2. Poor control over specific language usages.	2. Observation and check of daily oral and written expression.	2. Check pupil's test paper to identify the classes of usages missed by pupil. Check with text and course of study for grade emphasis. Emphasize individual drill on specific points of error. In teaching the correct forms, contrast them with the ones to be avoided. Sound is important in usage; supplement written exercises with oral drill. (See references 2 and 7.)
3. Poor language background.	3. Careless and inaccurate usage in oral and written expression; poor enunciation and pronunciation.	3. Corrective instruction is the only remedy here. Select a limited number of important usages and proceed as in No. 2 above. (See reference 7.)
4. Foreign language in the home.	4. Observed foreign accents in pronunciation. Use evidence of emphasis of two languages in home.	4. Use corrective instruction here. Follow suggestions in No. 2 above.
5. Poor general reading comprehension.	5. Erratic responses to test items; observation of reading ability in other subjects.	5. Drill on sentence and total meaning comprehension as required for general improvement in reading. (See reference 6.)
6. Low mental ability.	6. Difficulty in following directions, with erratic responses to items; difficulty in mastery of common usages; low mental age and IQ as shown by reliable mental test.	6. Follow general procedure outlined in Nos. 2 and 3 above. Have pupil prepare and memorize a key sentence for the troublesome usages.
7. Careless language habits.	7. Inaccurate and erratic response to test items; observed carelessness in informal expression.	7. Develop a self-critical attitude toward correct usages. Bring social and group pressure to bear favoring correct usages. Stimulate pupil to proofread all written expression before submitting it.
8. Confusion resulting from emphasis on formal rather than functional usages.	8. Inaccurate responses to items emphasized mainly through formal statements of rules.	8. Emphasize individual drills, stressing the establishing of definite habits of correct response to important usages. (See reference 2.)

<sup>11</sup> *Manual for Interpreting Iowa Language Abilities Tests*. World Book Co., Yonkers, N. Y., 1948. p. 24-31.

## PUNCTUATION

POSSIBLE CAUSES OF LOW TEST SCORES	ADDITIONAL EVIDENCE OF DEFICIENCY	SUGGESTED REMEDIAL TREATMENT
1. Lack of knowledge of the specific punctuation skills.	1. Check of test papers to determine types of skills missed in test; observation of daily written work.	1. Check the punctuation items missed in the test with the textbook and the local course of study. Use proofreading drills emphasizing the types of skills missed by pupil. Drill that calls for two different types of reaction in punctuation is desirable. Give some drill on inserting the correct punctuation where none is given. Give some drill on avoiding over-capitalization. Emphasize the need for critical self-editing of pupil's own or other copy. (See reference 3.)
2. Tendency to overpunctuate.	2. Analysis of test paper and daily written work for evidence of excessive punctuation marks, especially commas.	2. Use dictation and proofreading drills calling for the elimination of improper or excessive punctuation. (See references 2 and 8.)
3. Lack of a self-critical attitude toward own written work.	3. Careless punctuation in all daily written preparations; limited ability to note errors in punctuation in own or other written copy.	3. Emphasize importance of self-criticism of own daily written work. Drill on proofreading exercises designed to emphasize the correct use of capitals in the types of situations in which the pupil shows weakness.
4. Poor general reading comprehension.	4. Low scores on reading comprehension tests; slow or poor interpretation when reading in other subjects.	4. Drill on word meaning, sentence comprehension, and comprehension of total meaning of printed material in varied subject-matter fields.
5. Poor vision or hearing.	5. Observation; doctor's or nurse's examination.	5. Refer to doctor for medical attention. Move pupil to front of room or near teacher during instruction. Encourage pupil to make a special effort to write carefully and to make his punctuation marks distinctly.
6. Poorly developed sentence sense.	6. Low scores on Test 6, Intermediate Test; observation of daily usage.	6. Explain the various types of sentences and the relation of sentence structure to punctuation. Stress individual practice in writing sentences and punctuating them correctly. Use dictation and proofreading exercises calling for punctuation.
7. Carelessness in matters of form in written expression.	7. Observation and analysis of characteristics of handwriting and punctuation in daily work.	7. Stress continually the essentials of good form in written work. Insist that pupil edit and proofread his own papers before submitting them.



## LETTER FORM EXERCISE

Write as I dictate the following items in the proper form for a business letter:

Your home address is 140 Grand Avenue Boulevard.

Your home town is Bluffton, Pennsylvania.

The date is (*give present date*).

You are writing to Williams and Burke, Attorneys-at-law, whose address is Sprague Building, 10th and Ferry Sts., Long Beach, California.

Use the proper salutation (gentlemen) for a letter of this sort.

**Remedial exercises on sentence structure**

Exercises of the following sort will afford effective remedial drill for children having difficulties in sentence structure.

SENTENCE STRUCTURE EXERCISE <sup>12</sup>

In each of the following groups of sentences there are some statements which are not well expressed. Place a cross (X) before each such statement to show that it is not a good sentence.

I asked her the name of the book she was reading.

He was glad to leave, for he was tired and sick of the place, for he had made no friends.

She told me the names of her sister and kitty.

Mary had a good position. Which she left.

She is happier than I.

**Remedial drill on choice of words**

A great deal of difficulty is encountered by youthful (and adult) writers in the choice of words. In many cases it is a matter of choosing a word with a more exact shade of meaning, and in many cases it is a matter of knowing the form of a given word to use. This is particularly true in the field of adjectives and attributive nouns. Exercises such as the following, giving drill in choosing the correct compared form of certain adjectives, will be found helpful in cases in which specific diagnosis reveals this type of weakness.

<sup>12</sup> Adapted from Course of Study in Language, University Elementary School, University of Iowa, Iowa City.

EXERCISE IN CHOICE OF WORDS <sup>13</sup>

Read the following exercises and fill in the missing word needed to complete the meaning of the sentence. In each the missing word in an exercise is a form of the underlined word in the sentence.

1. New York is a large city. It is \_\_\_\_\_ than Chicago; in fact it is the \_\_\_\_\_ city in the United States.
2. I want to buy a good fountain pen. I want a \_\_\_\_\_ one than that. Show me the \_\_\_\_\_ one you have.
3. John is a mischievous boy. He is \_\_\_\_\_ than Bob. He is the \_\_\_\_\_ boy in school.
4. This is a bad storm, much \_\_\_\_\_ than the one we had last week, but not the \_\_\_\_\_ one this year.

#### 4 MEASUREMENT AND REMEDIATION IN SPELLING

##### Social and educational significance of spelling

The importance of correct spelling in the written communication of ideas is quite generally recognized. Applicants for positions have often failed to receive employment because of incorrect spelling of words in their letters of application. Business and social status is frequently determined to a large measure by a person's mastery or lack of mastery in this specific skill. Spelling, because of its social significance and its tool value in connection with later school progress, is so important that educators in general are unwilling to depend on the incidental teaching of it for the development of the required skill. Spelling is recognized as one of the subject fields in which the learning is specific. The child does not just learn spelling, but he learns to spell specific words. He may master a definite method of learning to spell, but the words he learns to spell are mastered as a result of a definite application of effort and attention.

*Objectives and measurable qualities in spelling.* While there are numerous other objectives for the teaching of spelling the primary aim is the development of mastery in the arrangement of the letters comprising the words most commonly used in writing. In addition to this aim there are at least two other important secondary aims: (1) the development of an attitude toward correctness and incor-

<sup>13</sup> *Ibid.*



rectness in spelling, and (2) the development of an effective method of learning how to master the spelling, meanings, and uses of new words. It is obviously impossible to provide objective measures for all of the specified outcomes of instruction. However, these outcomes furnish the best basis for the measurable qualities in the field. The ability to spell in list or in context those words that are most commonly needed in written expression can be measured quite satisfactorily by means of samplings of words taken from vocabulary lists of known social importance. The ability to recognize the correct or incorrect spellings of socially useful words is quite readily measured by means of proofreading tests. This same type of measure can be used within certain limits in the determination of the development of a "spelling conscience." It is doubtful whether there are any existing tests suitable for the measurement of the acquisition of new word meanings. This is also true of the development of good study habits in spelling. A close observation of the pupil's daily work in spelling probably affords the best check on his use of proper study habits. Good habits of work in spelling, as in other subjects, are usually revealed indirectly in the results.

Early in the consideration of the problems of measurement in spelling, two aspects of the pupil's accomplishment in this subject should be pointed out. In the first place, there is the problem of determining the pupil's present spelling ability. The second aspect of the problem concerns the measurement of progress. This is expressed in terms of the improvement the pupil makes, under instruction, in the mastery of the specific spelling vocabulary on which he is at work. Thus the determination of the child's spelling ability should be undertaken prior to study on the specific list of words. At the end of the teaching process a second measure is taken. This affords an indication of how much the child has progressed in his mastery of the selected spelling vocabulary.

### Systematic sampling of words

The introduction of scientific methods in education in recent years has resulted in many investigations into the scope and character of spelling lists. Studies such as those by Anderson, Ayres, Fitzgerald, O'Shea, Thorndike, and Horn seem to warrant the conclusion that approximately 4000 carefully selected words would be an appropriate number for the basic spelling list for elementary schools. Further-



more, these studies have proved of great value in selecting the word lists to be included in spelling texts, tests, and scales. It is quite obvious that the words most commonly used in the written language activities of adults and children should receive the major emphasis in a spelling course of study. To teach pupils to spell words that they will very rarely be called upon to spell either in or outside of the school is clearly a waste of time. Such words are best left to incidental learning or to the responsibility of each person as the need for their use arises.

## Construction of spelling tests

In the construction of spelling tests the following four problems require careful consideration.

*What words.* One of the first problems in the construction of a spelling test is that of selecting the words to be included. The values of spelling are almost entirely specific, and lie in the ability of the pupil to spell words that are actually used and are most certain to be used. It is important, therefore, that the content for a test should be sampled from those words that are and will be ultimately of maximal usefulness to the pupil. Theoretically, spelling lists may be taken from the writing vocabularies of children, the writing vocabularies of adults, or from words common to both.

Among the word lists that have been widely used in the construction of spelling tests is one by Anderson,<sup>14</sup> comprising the *Iowa Spelling Scales*, the Thorndike *Teacher's Word Book*,<sup>15</sup> and the Horn *Basic Writing Vocabulary*.<sup>16</sup> Anderson's list was one of the first to be based on an extensive word count. Thorndike's list contains ten thousand words that were found to occur most frequently in a count of several million words taken from many sources. Horn's list includes ten thousand words chosen from varied types of adult writing. The words are classified on the basis of frequency, and each word frequency is compared with that given in other vocabulary studies. This study took into account all previous spelling vocabularies, and

<sup>14</sup> W. N. Anderson, *Determination of a Spelling Vocabulary Based upon Written Correspondence*. University of Iowa Studies in Education, Vol. II, No. 1. University of Iowa, Iowa City, 1917.

<sup>15</sup> E. L. Thorndike, *The Teacher's Word Book*. Teachers College, Columbia University, New York, 1921.

<sup>16</sup> Ernest Horn, *A Basic Writing Vocabulary*. University of Iowa Monographs in Education. First Series, No. 4. University of Iowa, Iowa City, 1926.



as a result has greatly influenced the content of recent spelling tests. Only 3009 of the ten thousand words in this writing vocabulary were designated by the author as basic for elementary-school spelling lists.

The *Iowa Spelling Scales*,<sup>17</sup> representative of another source of information on what words to include in a spelling test, are based on the 2977 words found by Anderson to be most frequently used in written correspondence. *The Standard Elementary Spelling Scale* by Bixler<sup>18</sup> contains a total of 3679 words, some of which are of uncertain origin, secured by making a cross-check of twenty-seven spelling lists.

Teachers who are using spelling texts made up of word lists of unknown social importance will find such sources of great value in selecting valid content for their own tests. Words to comprise a spelling test should, of course, be among those comprising the list studied by the pupils. The most valid types of spelling words on which to test a pupil are also those words that have relatively high social usage. Thus a cross-check of the words common to the local spelling text and to a standardized spelling scale will reveal the high social-frequency words that the pupils have studied and will at the same time give the teacher a measure of the relative difficulty of the words from their values in the scale itself. Thus the teacher can construct his own valid test on words of known difficulty.

*How difficult words.* It is well known that some words are more difficult than others, i.e., some words are more frequently misspelled than others. If words are selected at random from any of the lists indicated above, some of them will be easy and some relatively difficult. Words for a test should be selected in terms of their known difficulty. The words in most spelling scales have been so classified by having the words spelled by large numbers of children and the relative difficulty of each word determined by the percentage of correct spellings of each word. The words to be included in the test for any grade should be adapted if possible to the ability of the group to be tested. Classes of average ability appear to respond best to words of approximately 50 per cent difficulty.<sup>19</sup> On the other hand,

<sup>17</sup> E. J. Ashbaugh, *The Iowa Spelling Scales*. Bureau of Educational Research and Service, State University of Iowa, Iowa City, 1945.

<sup>18</sup> H. H. Bixler, *The Standard Elementary Spelling Scale*. Turner E. Smith and Co., Atlanta, Ga., 1940.

<sup>19</sup> Walter W. Cook, *The Measurement of General Spelling Ability Involving Controlled Comparisons between Techniques*. University of Iowa Studies in Education, Vol. VI, No. 6. University of Iowa, Iowa City, 1932.

if the test is to be given over a wide spread of ability, words ranging from 14 to 86 per cent standard accuracy with a mean of 50 per cent tend to give a distribution more closely approximating the normal frequency curve, with the pupils grouped more closely around the mean. In general, it is probably safe to say that the words to be included in a test for any grade should be those on which there are from 40 to 70 per cent misspellings. Tests made up of such words will give a reliable measure of spelling ability, since the words will not be so easy that there will be many perfect scores or so difficult that there will be many low scores.

*How many words.* The purpose the test is to serve will determine the number of words to use. For survey purposes a list of 25 words will probably be sufficient to determine the status of spelling efficiency for a school system. To be sure, the ability to spell one word is separate and distinct from the ability to spell other words. It would seem necessary, therefore, to subject a pupil to several hundred words in order to secure a reliable measure of his ability to spell the most commonly used words. However, the procedure of sampling applies to the testing of spelling as in all other testing. While 25 words is possibly a sufficient number for survey purposes, a larger number is needed to reveal the spelling ability of individual pupils. On the whole it appears that a minimum of 100 words should probably be used for individual testing purposes in spelling. Possibly 50 words are not too many to use for the measurement of general class accomplishment.

*How given.* The question of the form in which spelling words should be presented for testing purposes has called forth much debate in the past. It has also been subjected to experimental study with results that are not too conclusive when considered in the light of practical classroom procedures. Horn summarized the evidence on this question as follows:<sup>20</sup>

Written tests are to be preferred to oral tests. . . Recall tests are superior to and more difficult than recognition tests. The evidence indicates that the most valid and economical test is the modified sentence recall form, in which the person giving the test pronounces each word, uses it in an oral sentence, and pronounces it again. The word is then written by the students.

<sup>20</sup> Ernest Horn, "Spelling," *Encyclopedia of Educational Research*, Revised edition. Macmillan Co., New York, 1950. p. 1259.



## Diagnosis and remediation of spelling disabilities

Spelling tests and scales afford valuable sources of material that may be used to determine both the pupil's present *status* in spelling and his *growth* in accomplishment as a result of a period of instruction. If scales based on a sound philosophy of subject-matter content are used, they provide the most effective materials for the identification of the spelling difficulties of individual pupils. Samplings from scales used as tests give the teacher an objective basis for the study of these personal difficulties through the accumulation of individual lists of words that are sources of trouble.

To a large extent remedial procedures in spelling may be undertaken directly in connection with teaching. The words misspelled by pupils in their spelling lessons and tests are obviously the words to which they should give special attention. Each pupil should be encouraged to keep an individual list of such words and should be stimulated to master them. Occasional spelling periods should be put aside for studying and testing these individual lists. If such lists are properly utilized, each pupil will come to regard his "demon" list as an effective means for eliminating spelling deficiencies.

Written work in all subjects should be carefully checked for spelling errors. A list of such misspellings should be kept by every pupil, and he should realize that he is to be held responsible for the mastery of these troublesome words. The important thing is that the learning situation be so manipulated that the pupil will want to learn to spell and to feel the need for learning the meaning and spelling of words that are pertinent to his written work.

### Individual pupil diagnosis

The discovery from the results of a spelling test that a pupil is below the norm in spelling ability may be of considerable value, but it falls far short of its real function unless it reveals to the pupil the particular weaknesses that resulted in his low score. The following items of information procurable through observation and measurement are invaluable in diagnosing individual pupil disabilities and should be used as much as possible in connection with the analysis of the pupils' spelling habits: (1) intelligence quotients, (2) spelling marks, (3) reading marks, (4) writing marks, (5) attendance data, (6) visual-defects data, (7) auditory-defects data, (8) speech data,

(9) general health data, (10) personality characteristics—industry, aggressiveness, independence, attentiveness, exactness.

In Tidyman and Butterfield the following procedure in diagnosing and treating problem cases in spelling is suggested:<sup>21</sup>

1. Give a standard spelling test to discover the amount of deficiency. Compare with achievement in other subjects.
2. Give an intelligence test to discover general mental capacity.
3. Test for defects of hearing and vision.
4. Give reading test.
5. Give test of spelling consciousness to show whether mistakes are due to carelessness or ignorance of the word.
6. Collect misspellings from spelling tests and written work, and classify them according to types of errors.
7. Get as much information as possible about the pupil's pedagogical history, especially methods of beginning reading; knowledge of meanings of words; knowledge of phonics; pronunciation and articulation; motor coordination in writing; and emotional attitude toward spelling.
8. From above, assemble probable causes of difficulty in spelling, and adopt appropriate remedial measures, such as the following:
  - (a) Systematic word study. Early training may have been inadequate.
  - (b) Exercises in visualization.
  - (c) Drill upon particular types of spelling errors.
  - (d) Phonics drills.
  - (e) Removal of physical defects.
  - (f) Develop confidence through successful effort.

## Remedial work in spelling

Poor spelling is due to faulty or inadequately formed associations. Basically, all spellers, good or bad, learn in the same way—through association. The main difference between the able and the poor speller lies in the study technique used, his personality characteristics, and the emphasis he gives to the subject.

Many investigators of spelling disabilities have abandoned the procedure of deducing the causes of spelling difficulties from an analysis of errors and are now devoting their time and energies to

<sup>21</sup> Willard F. Tidyman and Marguerite Butterfield, *Teaching the Language Arts*. McGraw-Hill Book Co., Inc., New York, 1951. p. 359, quoted from William H. Burton, ed., *Supervision of Elementary Subjects*. D. Appleton & Co., New York, 1929. p. 121-22.



TABLE 32. Remedial chart: spelling

POSSIBLE CAUSES OF LOW TEST SCORES	ADDITIONAL EVIDENCE OF DEFICIENCY	SUGGESTED REMEDIAL TREATMENT
1. Lack of experience with the testing technique.	1. Low score on test contrasted with high score when words are given on dictation test.	1. Drills on choosing correct spellings from lists of errors of same word; choosing correctly spelled words from long lists, some of which are spelled correctly, some incorrectly; proofreading own written work carefully.
2. Instructional emphasis on different or wrong vocabulary.	2. Low score on test in contrast with good record for spelling in daily work.	2. Check the words not taught in your spelling course of study with lists of known social utility. (See references 1 and 5.)
3. Failure to develop a critical attitude toward good spelling.	3. Noticeable indifference to spelling errors in daily written work.	3. Emphasize critical proofreading of own work. Drill on choosing correct spellings in lists of words. Check pupil's certainty of his judgment of correctness of spelling.
4. Lack of teaching emphasis on individual's own spelling difficulties.	4. Observation of pupil's misspellings in daily written work.	4. Have pupils keep lists of words misspelled in daily work and use these as basis for individual study. Work on pupil's own errors. See what transfer is made to other written work.
5. Specific learning difficulties. a. Faulty pronunciation of words by the teacher.	5. a. Observation of speech habits; informal pronunciation tests based on spelling vocabulary.	5. a. Look up word in dictionary. Pronounce it distinctly for pupil. Have him repeat it while looking at the written form of word to associate sight with correct sound.
b. Difficulties in seeing or hearing.	b. Observation; doctor's or nurse's examination.	b. Refer to doctor for medical advice. Move pupil to front of room near window and blackboard. Stand near him during all tests and spelling exercises. Make special effort to speak and write clearly.

<i>c.</i> Limited power to visualize or "see" word forms.	<i>c.</i> Observation test. Try visualization test, as trying to see in mind's eye a three-inch block painted red. Ask detailed questions about number of faces, number of planes to cut it into one-inch cubes, number of small cubes, etc.	<i>c.</i> Emphasize the practice of looking at the word, closing the eyes, and attempting to recall the word, as part of every spelling study period.
<i>d.</i> Failure to associate sounds of letters and syllables with spelling of words.	<i>d.</i> Individual interview; analysis of spelling errors in tests and daily work.	<i>d.</i> Go over words with child before he studies them in spelling. Teach him to analyze words himself.
<i>e.</i> Tendency to transpose, add, or omit letters.	<i>e.</i> Analysis of spelling test papers; observation of daily work; pronunciation tests.	<i>e.</i> Emphasize visual recall of words. Have pupil practice writing the words, exaggerating the formation of letters. Underline individual hard spots.
<i>f.</i> Tendency to spell unphonetic words phonetically.	<i>f.</i> Note types of errors made in spelling tests, especially insertion or omission of letters.	<i>f.</i> Show that all words are not spelled as they sound. Spelling of each word must be learned individually. Emphasize steps in learning to spell. See <i>h</i> below.
<i>g.</i> Difficulties in writing; letter formation.	<i>g.</i> Observation of daily written work and spelling papers. Check with handwriting scales, as Ayres or Freeman Charts.	<i>g.</i> Practice in difficult letter formations and combinations. Emphasize need to avoid confusing letter forms, as <i>i</i> , <i>e</i> , <i>r</i> , <i>t</i> . (See reference 4.)
<i>h.</i> Failure to master a method of learning to spell.	<i>h.</i> Low scores on daily spelling tests; observation of pupil method of study in spelling; test on steps in learning to spell.	<i>h.</i> Check child's method of learning in spelling. Teach child steps in learning to spell until he uses them. Suggested steps: (1) Look at word. (2) Listen while teacher pronounces it. (3) Pronounce it by syllables. (4) Use it in a sentence. (5) Close eyes and visualize it. (6) Write it. (7) Close eyes and recall. (8) Write word. Repeat steps as necessary. (See reference 2.)



studying the work habits of pupils by means of careful observation and tests. An excellent summary of many causes of spelling deficiency together with suggested remedial treatment is given in the accompanying chart adapted from a current language test manual.<sup>22</sup>

## 5 MEASUREMENT AND REMEDIATION IN HANDWRITING

In spite of increased availability, popularity, and use of mechanical means for writing, both in school and out, there is little reason to believe that handwriting will be displaced as the major means of written communication. If it is to serve as an adequate tool in social and business communication, handwriting must be easily read, neat and pleasing in appearance, and of such form that it can be produced under normal conditions with a fair degree of speed. A functional program in handwriting, according to a discussion of the subject, is one which "(1) defines competency in terms of standards acceptable in the social and business correspondence of adults; (2) encourages individuality of style; (3) emphasizes legibility, appearance, and ease of writing; (4) eliminates formal drills and limits practice to meeting immediate, recognized needs; (5) relates handwriting to written composition; (6) favors a natural arm-hand-finger movement, adapted to age and maturity; and (7) permits the use of handwriting materials commonly used in the home and the business world."<sup>23</sup>

*Objectives and measurable qualities in handwriting.* A concise summary of the objectives of instruction in handwriting is given below. While this statement is not especially new or recent it nevertheless provides a very useful basis for the identification of the measurable objectives in this important skill of expression.<sup>24</sup>

### OUTCOMES OF HANDWRITING INSTRUCTION

1. To develop sufficient skill to enable pupils to write easily, legibly, and rapidly enough to meet present needs and social requirements.
2. To equip the child with methods of work so that he will attack his writing problems intelligently.

<sup>22</sup> *Manual for Interpreting Iowa Language Abilities Tests*. World Book Co., Yonkers, N. Y., 1948. p. 24-25.

<sup>23</sup> Tidyman and Butterfield, *op. cit.* p. 362.

<sup>24</sup> "Handwriting." *The Nation at Work on the Public School Curriculum*. Fourth Yearbook of the Department of Superintendence. National Education Association, Washington, D. C., 1926. p. 113-14.

3. To diagnose individual writing difficulties.
4. To aid the child to recognize and make use of his peculiar learning capacities.
5. To provide experiences which will tend to develop in the child more power to direct his own practice, and more ability to judge whether or not he is succeeding in that practice.
6. To provide the means for each individual to progress at his best rate.
7. To develop an appreciation of the relationship between correct body adjustment and an efficient writing production.
8. To secure acceptable and customary arrangement and form for written work (margins, spacing, etc.).
9. To develop a social urge to use the skill attained in all writing situations.
10. To train pupils to be able, at the end of the sixth grade, to write quality 60 (Ayres Scale) or better, and at the rate of 70 letters per minute or better.

Writing involves a very exact type of visual-muscular coordination, which must be developed to a high degree if the product is to possess legibility, speed of production, and æsthetic quality. Some difficulty has been encountered in the measurement of certain of the elements of good writing, particularly from the point of view of analysis and diagnosis. The available writing scales, however, have done much to establish for the pupil and teacher rather definite standards or ideas of what constitutes an acceptable product as well as to make both more and more sensitive to handwriting faults and needs.

The measurement of handwriting quality in its refined form is concerned with two factors: (1) quality, or degree of legibility, and (2) speed, or the quantity of writing produced in a given unit of time.

*Quality.* The quality of handwriting is usually determined by comparing a sample of handwriting with specimens in a standard scale. While this method of evaluating handwriting specimens is somewhat subjective, experience shows that considerable skill and objectivity can be developed through training in the use of such scales. At one time the measurement of handwriting simply involved the comparison of the script produced with the copybook sample. This resulted in overemphasis on the shape and shading of letters and in the formation of beautifully engraved lines. Rate and quality were not the objectives of writing instruction or of measurement under those conditions.

The essentials of quality in writing are measurable within reason-



able limits. A number of scales have been developed for use in measuring quality, but they differ greatly in the type of copy used, in the number of elements of quality measured, and in the numerical designation of each quality difference. Therefore, it is difficult to compare the results secured from the use of one scale with those secured from another.

West and Freeman <sup>25</sup> proposed the quality and rate norms for the *Ayres Scale* as shown in Table 33. These norms for quality are considered by many to be sufficiently high, although the values assigned for the various grades are based upon the median performance of many school children. The norms proposed by West and Freeman do not agree exactly with those accompanying the *Ayres Scale*. Perhaps the Ayres data may be considered as norms, while the quality and rate values proposed by Freeman may more nearly represent standards.

TABLE 33. Handwriting quality and speed standards

Grades	2	3	4	5	6	7	8
Quality on Ayres Scale.....	44	47	50	55	59	64	70
Rate in Letters per Minute....	36	48	56	65	72	80	90

Koos investigated the quality of adult handwriting and also the opinions of adults concerning a satisfactory quality of handwriting. He reached the following conclusions on the basis of his findings: <sup>26</sup>

The fact that some (pupils) will go into pursuits demanding a quality better than 60 should not be offered as a justification for requiring all pupils to attain that better quality. . . Since all should be required to learn to write as well as 60 for purely social use, to train pupils to write this quality is the task of general education; to teach some who are going into commercial or other vocations requiring a higher quality . . . to write this better quality is the task not of general but of vocational educa-

<sup>25</sup> Paul V. West and Frank N. Freeman, "Handwriting." *Encyclopedia of Educational Research*, Revised edition. Macmillan Co., New York, 1950. p. 524-29.

<sup>26</sup> L. V. Koos, "The Determination of Ultimate Standards of Quality in Handwriting for the Public Schools." *Elementary School Journal*, 18:423-46; February 1918.

tion... In the light of these facts it is difficult to see why... a pupil should be required to spend the time necessary to learn to write better than the quality of 60. There is even considerable justification for setting the ultimate standard at 50.

*Rate.* The rate at which pupils write is of considerable importance. The person who is able to write more rapidly than others and with approximately the same quality has an obvious advantage in the field of written expression, provided, of course, that ideas come to him as rapidly as he is able to transcribe them. The measurement of rate in writing is much less difficult than the measurement of quality. Rate of writing can be measured most conveniently by asking pupils to write, within carefully controlled time limits, selections from standardized copy. If the pupils all write from the same selection and if they have all thoroughly memorized it, the number of letters each pupil writes in the time allotted can easily be computed as the pupil's rate score. Table 33 indicates that pupils at the end of the second grade should write at the rate of 36 letters per minute and increase their speed to 48 letters per minute by the end of the third grade and to a rate of 90 letters per minute by the end of the eighth grade.

### Measurement of handwriting ability

Measurement of the quality of handwriting and the rate at which it is produced is accomplished by the evaluation of handwriting specimens secured under standard conditions. Accordingly, the first step in the process of measuring handwriting is that of securing these specimens under controlled conditions.

*Securing handwriting specimens.* Three factors appear to affect the conditions under which handwriting specimens for scaling are secured. In the first place, the character of the copy which the pupils are called upon to write may significantly influence their reactions. It is usually better, at least in the lower grades, to make use of some simple sentence or paragraph which the children have memorized in previous connections, such as "Mary had a little lamb," or some other equally familiar nursery rhyme. The sentence, "The quick brown fox jumps over the lazy dog" has been used on numerous occasions. The chief merit of this sentence lies in the fact that it contains all of the letters of the alphabet. Whatever sample is used



should be simple and easily understood, so that the children will not be unduly affected by spelling and vocabulary difficulties. To guard against lapses in memory, it is good practice to write the copy on the blackboard two or three days in advance of the tests where it can be easily seen and studied both before and during the collection of the writing specimens.

The instructions which are given to the pupils may also influence the quality and rate of their writing. Therefore, care should be exercised to use very precise directions. The use of this statement in the instructions to the children is recommended: "Write as well as you can and as rapidly as you can."

The time allowance for the writing of the specimens is a third factor which must be considered in the collection of writing specimens. In the standardization of his scale Ayres used the first four sentences of Lincoln's Gettysburg Address and allowed each child two minutes in which to copy as much of this material as possible. Since that time it has become a rather typical practice to allow a two-minute period for the writing of such samples.

The teacher who is inexperienced in the administration of such a test may find the following directions helpful. When the children are all ready, having been provided with paper, and pen and ink or pencil, depending on the grade and the course of study, they should be given a few simple directions. The following are suggested: "Write as well as you can at your usual speed, using the following copy: 'Mary had a little lamb' (or some other selected copy). Write the copy over and over until I give the signal 'Stop.' When I say 'Stop' you should stop even though you are in the middle of a letter." After these directions have been given the teacher may say, "All in position. Ready. Begin." At the expiration of two minutes the signal "Stop" should be given and the pupils asked to place their pens or pencils on their desks.

*Securing rate scores.* Rate of handwriting is expressed in terms of the number of letters written per minute. This is determined by counting the total number of letters written by each pupil and dividing this by the number of minutes the pupils were allowed to write.

*Securing quality scores.* The quality of the handwriting specimen being measured is determined by moving it along the scale until a specimen is found that closely matches it in quality. The quality value of the scale sample is then assigned as the quality of the

sample of the pupil's handwriting. As the scorer gains experience intermediate values may be estimated.

*Accuracy in measurement of handwriting.* Skill in the evaluation of handwriting specimens requires a thorough understanding of the scale to be used and considerable training in its use. It is desirable, therefore, for the teacher, prior to any attempt to use the scale for the measurement of handwriting quality, to study carefully the scale itself, the directions for its use, the norms, and the specific functions which the particular scale is expected to perform. The accurate and reasonably objective rating of handwriting samples on a scale requires considerable skill, which experience shows can be developed through practice. For this purpose standard sets of writing samples of known quality are very useful.

## Merit scales

Handwriting scales may be divided into two groups: (1) general merit scales and (2) analytical and diagnostic charts and scales. The choice of a scale depends on the purpose it is to serve.

The *Thorndike Scale* was the first writing scale to be devised. This scale is designed for Grades 5 to 8 inclusive and consists of a series of specimens of handwriting so arranged that they increase in order of merit from a quality of 4 units above zero to one of 18. Its purpose is to aid teachers in grading handwriting for "general merit" on the basis of three characteristics: beauty, legibility, and character.

*The Ayres Handwriting Scale*,<sup>27</sup> the next scale to be devised, was standardized on the basis of legibility. Legibility was determined by the speed and ease with which the samples of handwriting were read by a number of trained and competent judges. The *Gettysburg Edition*, now in general use, contains only one style of handwriting—the accepted moderate-slant style.

*The American Handwriting Scale* developed by West is one of the most recent and comprehensive of the general merit scales. Among a number of distinctive features are at least two that deserve special mention: (1) A separate scale is provided for each grade from 2 to 8;

<sup>27</sup> Leonard P. Ayres, *A Scale for Measuring the Quality of Handwriting of School Children*. Bulletin No. 113. Division of Education, Russell Sage Foundation, New York, 1912.



(2) The samples have been scaled for both quality and rate, the poorer samples being written at a slower rate and the better samples being the ones written at a more rapid rate. The existence of the separate scales for Grades 2 to 8 inclusive permits a somewhat more exact evaluation of quality of writing in its relation to grade location.

*The Conard Manuscript Writing Standards*<sup>28</sup> are composed of two scales for the rating of manuscript writing. Two separate scales, one for the rating of pencil forms and the other for the rating of pen work, are available.

*Diagnosis and remediation of handwriting.* Instruments for the identification of specific faults in handwriting are of two general types: (1) analytical scales and (2) score cards.

*The Freeman Chart for Diagnosing Faults in Handwriting* is virtually five scales in one. Each scale is designed to reveal whether the pupil's writing specimen violates one or more of the five essential characteristics of good handwriting. These traits are: (1) uniformity of slant, (2) uniformity of alignment, (3) quality of line, (4) letter formation, and (5) spacing. Each scale shows three levels of quality for the trait with which it deals—excellent, mediocre, and poor. This scale is valuable because it enables both teacher and pupil to discover the handwriting weaknesses that are in need of special treatment.

## Diagnosis by analysis

Improvement in handwriting instruction depends to a large degree on the teacher's knowledge of the elements that make for quality in the product, and the use of instruments that are adequate to reveal significant differences in quality. Inferior products of handwriting instruction may be due to lack of skill or mastery in many different phases of the writing act. Freeman's *Chart for Diagnosing Faults in Handwriting* meets this need for securing separate measures of the several aspects of handwriting performance. This scale may be used to measure the whole class, but it is most effective when used to diagnose the writing of pupils who rank conspicuously below the grade norm as revealed through use of some general merit scale.

The following list of handwriting defects and their causes should be useful to the classroom teacher.

<sup>28</sup> Edith U. Conard, "Manuscript Writing Standards." *Teachers College Record*, 30:669-80; April 1929.

## ANALYSIS OF DEFECTS IN HANDWRITING AND THEIR CAUSES

<i>Defect</i>	<i>Causes</i>
1. Too much slant . . . . .	(1) Writing arm too near body (2) Thumb too stiff (3) Point of nib too far from fingers (4) Paper in wrong position (5) Stroke in wrong direction
2. Writing too straight . . .	(1) Arm too far from body (2) Fingers too near nib (3) Index finger alone guiding pen (4) Incorrect position of paper
3. Writing too heavy . . . . .	(1) Index finger pressing too heavily (2) Using wrong pen (3) Penholder too small diameter
4. Writing too light . . . . .	(1) Pen held too obliquely or too straight (2) Eyelet of pen turned side (3) Penholder too large diameter
5. Writing too angular . . . .	(1) Thumb too stiff (2) Penholder too lightly held (3) Movement too slow
6. Writing too irregular . . .	(1) Lack of freedom of movement (2) Movement of hand too slow (3) Pen gripping (4) Incorrect or uncomfortable position
7. Spacing too wide . . . . .	(1) Pen progresses too fast to right (2) Too much lateral movement.

**Physical conditions and materials**

Prominent among the physical factors affecting the pupil's handwriting is his desk. The pupil's desk should be adjusted to his height so that when he is seated normally his thigh is at right angles to the lower part of his leg and his feet are flat on the floor. In accordance with most modern methods of writing, the pupil's body, when he is writing, should face the middle of the desk squarely and bend slightly forward at the hips. Both forearms should be well up on the desk, the left holding the paper, the right wrist raised and inclined slightly to the right. It is necessary that the pupil be taught to move the paper upward and to the left as the writing progresses. The shifting is done with the left hand, while the right arm is held in the correct



position. There is some difference of opinion about the best position of the writing arm. It is generally agreed, however, that the writing hand should be supported on the third and fourth fingers and that the wrist should not be tilted more than 45 degrees. The forearm of the right hand should be perpendicular to the line of writing. The pen should be grasped lightly and in such a way that the forefinger is below the thumb and at least one inch above the point of the pen.

Experiment and observation show that modern writing is a combination of whole arm, forearm, wrist, and finger movements. It is not possible or even desirable to eliminate finger movement entirely, even in so-called "muscular movement writing."

### Improvement of psychological conditions

Next in importance in preparing the way for effective mastery of writing faults is the provision of desirable psychological conditions. The establishment of a desire for improvement on the part of the pupil is essential. One plan that has been proved to be quite effective involves the pupils' use of handwriting scales for the appraisal of their own writing. A copy of some good general merit scale should be conveniently placed in every classroom to encourage and train pupils in its use as a means of facilitating comparisons and evaluation of personal products.

Another means of motivation is the exemption from further penmanship drill of all pupils who have attained an acceptable standard of speed and quality. The standard of 60 for speed and quality is the one generally accepted. Evidence seems to indicate that from 50 to 75 per cent of the pupils in the upper grades can easily reach this standard. If these pupils are exempt from further drill, the teacher is able to devote more time to those who have failed to meet the standard.

For improving the rate of slow writers, the writing of some simple sentence such as one of those given on page 451 is recommended. If the sentence, "The quick brown fox jumps over the lazy dog," is used, the following table indicates the number of times the sentence should be written in the specified time limits. Instruction in the making of different letters may be required by some children. Pupils are greatly helped by special practice on the letters and strokes that have given them trouble as a means of enabling them to attain accepted standards of speed and quality.

TABLE 34. Tentative norms for writing a practice sentence

Grade	Number Times Written	Time in Minutes and Seconds
8	11	4 00
7	8	3 00
6	6	3 00
5	5	2 45
4	4	2 30
3	3	2 10
2	2	2 00

### Handedness as a factor in writing

In addition to the physical and psychological conditions discussed in the preceding paragraphs, there is the very important factor of handedness in the pupil. The general considerations of method and remedial procedures in handwriting appear to assume right-handedness in the child. Yet left-handedness is common enough in the classroom to represent a significant problem to the teacher, and one worthy of some consideration here. Naturally enough the pursuit of methods of instruction and remedy suitable for the right-handed pupil results in the formation of atrocious writing habits for the left-handed pupil. Any attempts to force him to conform to common right-handed practices usually forces him to write backwards, i.e., toward the left. In order to correct for the resultant reversal of the image, the pupil frequently twists his left wrist around in such a way that the pencil or pen point is directed toward him, with the result that he works awkwardly and under a most severe muscular maladjustment. For these and for other reasons that appear to be related to the speech and language functions, the teacher should probably not attempt to change over the left-handed pupil. It is almost certainly better to accept the tendency to left-handedness which is well developed by the time the child enters the first grade and to aid him in making the best possible adjustments and adaptations in his mastery of handwriting than it is to run the risk of confusing him and possibly causing serious emotional disturbances at a later time. There is little or no evidence that the child at birth has any predispositions to use one hand rather than the other. Since he lives in a predominantly right-handed world in which it is easier to



conform than not to conform, parents may profitably give some serious attention to the problem during the child's early formative years. The proper time to affect the child's handedness without danger of harmful reactions would seem to be in the period from his first active moment until he reaches school age.

### Topics for Discussion

1. What are the major situations in life in which language is used?
2. Evaluate the relative demands made by life situations on the oral and written aspects of language.
3. From the standpoint of classroom emphasis, which should receive more attention, oral or written language?
4. According to Travis, what are the major causes of oral language disabilities?
5. Discuss the problems of measurement of oral language abilities.
6. How is ability in written composition measured?
7. Discuss the status of analytical testing of written language abilities.
8. Discuss some remedial drill materials of value in language instruction.
9. What appears to be the most acceptable fundamental assumption upon which the spelling vocabulary suitable for elementary-school instruction should be based?
10. Show how a spelling test can be made from a standard spelling scale.
11. How can a spelling test made up of socially useful words be validated for use in a classroom in which a textbook in spelling based on a vocabulary of unknown social significance is in use?
12. What range of difficulty in words would you select for the purpose of measuring a class of unusually poor spelling ability?
13. Discuss the pupil work habits that have diagnostic significance in the field of spelling.
14. Which of the objectives or outcomes of instruction in handwriting are most defensible from a social utility point of view?
15. What is the relationship between handwriting speed and quality?
16. What place have standards in the evaluation of handwriting?
17. Describe some of the methods of diagnosing handwriting ability.
18. Present your reactions to the use of manuscript writing in the primary grades.
19. Discuss the implications of parental responsibility for left-handedness and the relation of handedness to handwriting, language, and speech.

## Selected References

- ASHBAUGH, E. J. *Iowa Spelling Scales*. Iowa City: Bureau of Educational Research and Service, State University of Iowa, 1945.
- BETTS, EMMETT A. "Guidance in the Critical Interpretation of Language." *Elementary English*, 27:9-18 ff.; January 1950.
- BOYNTON, MARCIA. "Inclusion of 'None of These' Makes Spelling Items More Difficult." *Educational and Psychological Measurement*, 10:431-32; Autumn 1950.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 294-317, 323-32, 541-43.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p.100-40.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938. p. 72-79.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. p. 218-47.
- CALLEWAERT, H. "A Rational Technique of Handwriting." *Journal of Educational Research*, 41:1-12; September 1947.
- COOK, WALTER W. "Evaluation in the Language-Arts Program." *Teaching Language in the Elementary School*. Forty-Third Yearbook of the National Society for the Study of Education, Part II. Chicago: Department of Education, University of Chicago, 1944. Chapter 9.
- DAWSON, MILDRED. *Language Teaching in Grades One and Two*. Yonkers, N. Y.: World Book Co., 1949.
- DIEDERICH, PAUL B. "The Measurement of Skill in Writing." *School Review*, 54:584-92; December 1946.
- DIEDERICH, PAUL B. "Teaching English with Test Exercises." *School Review*, 55:80-86; February 1947.
- DOLCH, EDWARD W. *Better Spelling*. Champaign, Ill.: Garrard Press, 1942.
- FORAN, THOMAS G. *The Psychology and Teaching of Spelling*. Washington, D. C.: Catholic Education Press, 1934.
- FREEMAN, FRANK N. "Contributions of Research to Special Methods: Handwriting." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1938. Chapter 6.
- FREEMAN, FRANK N., AND DOUGHERTY, MARY L. *How To Teach Handwriting*. Boston: Houghton Mifflin Co., 1923.
- GREENE, HARRY A. *A Criterion for the Course of Study in the Mechanics*



- of Written Composition*. University of Iowa Studies in Education, Vol. VIII, No. 4. Iowa City: University of Iowa, November 4, 1933.
- GREENE, HARRY A. "Contributions of Research to Special Methods: English Usage." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1938. Chapter 9.
- GREENE, HARRY A. "English—Language, Grammar, and Composition." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 383-96.
- GREENE, HARRY A., AND GRAY, WILLIAM S. "The Measurement of Understanding in the Language Arts." *The Measurement of Understanding*. Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1946. p. 176-89.
- HARRIS, CHESTER W. "Prediction of the Difficulty Index of Objective-Type Spelling Items." *Educational and Psychological Measurement*, 7:319-25; Summer 1947.
- HEFFERNAN, HELEN W. "Readiness for Oral and Written Language." *Elementary English*, 27:247-53; April 1950.
- HILDRETH, GERTRUDE H. "Evaluation of Spelling Word Lists and Vocabulary Studies." *Elementary School Journal*, 51:254-65; January 1951.
- HORN, ERNEST. *A Basic Writing Vocabulary: 10,000 Words Most Commonly Used in Writing*. University of Iowa Monographs in Education, First Series, No. 4. Iowa City: State University of Iowa, 1926.
- HORN, ERNEST. "Contributions of Research to Special Methods: Spelling." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1938. Chapter 8.
- HORN, ERNEST. "Spelling." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 1247-64.
- Iowa Elementary Teachers Handbook, Vol. 3. *Spelling and Handwriting*. Des Moines: Iowa State Department of Public Instruction, 1943.
- Iowa Elementary Teachers Handbook, Vol. 4. *Oral and Written Language*. Des Moines: Iowa State Department of Public Instruction, 1944.
- JOHNSON, LESLIE W. "One Hundred Words Most Often Misspelled by Children in the Elementary Grades." *Journal of Educational Research*, 44:154-55; October 1950.
- JOHNSON, WENDELL, AND OTHERS. *Speech Handicapped School Children*. New York: Harper and Brothers, 1948.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. p. 117-41, 144-52, 156-67.

- LESTER, JOHN A., AND LINDQUIST, E. F. "Examinations in English." *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Co., 1936. p. 410-37.
- LEWRY, MARION E. "Writing Vocabulary for Grade I." *Elementary School Journal*, 48:88-90; October 1947.
- MARCKWARDT, ALBERT H., AND WALCOTT, FRED G. *Facts about Current English Usage*. National Council of Teachers of English, English Monograph No. 7. New York: D. Appleton-Century Co., Inc., 1938.
- McKEE, PAUL. "An Adequate Program in the Language Arts." *Teaching Language in the Elementary School*. Forty-Third Yearbook of the National Society for the Study of Education, Part II. Chicago: Department of Education, University of Chicago, 1944. Chapter 2.
- McKEE, PAUL. *Language in the Elementary School*. Boston: Houghton Mifflin Co., 1939.
- MULGRAVE, DOROTHY I. *Speech for the Classroom Teacher*. Revised edition. New York: Prentice-Hall, Inc., 1946.
- SPACHE, GEORGE. "Spelling Disability Correlates I—Factors Probably Causal in Spelling Disability." *Journal of Educational Research*, 34:561-86; April 1941.
- STRICKLAND, RUTH G. *The Language Arts in the Elementary School*. Boston: D. C. Heath and Co., 1951.
- THORNDIKE, EDWARD L. *A Teacher's Word Book*. Revised edition. New York: Teachers College, Columbia University, 1931.
- TIDYMAN, WILLARD F., AND BUTTERFIELD, MARGUERITE. *Teaching the Language Arts*. New York: McGraw-Hill Book Co., Inc., 1951.
- TIEGS, ERNEST W. *The Management of Learning in the Elementary Schools*. New York: Longmans, Green and Co., 1937. Chapters 7-8.
- TRAVIS, LEE E. "Diagnosis in Speech." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1935. Chapter 19.
- TRAXLER, ARTHUR E. *The Use of Test Results in Diagnosis and Instruction in the Tool Subjects*. Revised. Educational Records Bulletin No. 18. New York: Educational Records Bureau, January 1937. p. 31-43, 62-73.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapters 11-13.
- WEITZMAN, ELLIS, AND McNAMARA, WALTER J. *Constructing Classroom Examinations*. Chicago: Science Research Associates, 1949. p. 54-63.
- WEST, PAUL V., AND FREEMAN, FRANK N. "Handwriting." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 524-29.
- WOOD, BEN D., AND HAEFNER, RALPH. *Measuring and Guiding Individual Growth*. New York: Silver Burdett Co., 1948. p. 291-98.



## ***Measuring and Evaluating in the Social Studies***

THIS CHAPTER presents a summary of the following points in the improvement of measurement and instruction in the social studies:

- A. Nature and organization of the social studies.
- B. Objectives and outcomes of the social studies.
- C. Kinds of tests in the social studies.
- D. Standardized tests in history, civics, and geography.
- E. Classroom testing and evaluating in the social studies.

The social studies deal primarily with past and current problems of human relationships and with the interactions of human beings as they associate with one another in varied political, economic, and social activities. Such school subjects as history, geography, civics, sociology, and economics are included in this area. Carr and his colleagues stated that the social sciences "are those bodies of scholarly materials which deal with human relationships," and that the social studies "are those portions of the social sciences which have been selected for instructional purposes."<sup>1</sup>

Two other terms in the area of the social sciences have recently come into use. Social learning is conceived by Moffatt and Howell to be broader than the social studies and to include "the social growth and development of the child as achieved through his total experiences," whereas they indicated that social education, sometimes

<sup>1</sup> Edwin R. Carr and others, "Social Studies." *Encyclopedia of Educational Research*, Revised edition. Macmillan Co., New York, 1950. p. 1214.

used as a synonym for social studies, "applies to all those activities that contribute to the child's social learning."<sup>2</sup> The broadened concept of the social studies embodied in these statements is reflected in portions of this chapter.

## 1 SCOPE OF SOCIAL STUDIES

The formulation of definite objectives in the social studies is a major problem, for research techniques useful in the establishment of objectives in such skill areas as arithmetic, reading, language, and spelling are difficult to apply in this area. There are, in fact, probably no scientifically established objectives for the social studies, which remain, in contrast to the areas emphasizing skills, a field in which content occupies a central position. This is still true even though modern social studies instruction places much more emphasis upon social skills than do more traditional methods.

### Objectives of the social studies

The rather detailed list given below specifies the understandings, attitudes, and skills or abilities that social studies instruction should develop in pupils.<sup>3</sup>

#### 1. Understandings

- a. Of the democratic faith and its meaning for human welfare and happiness
- b. Of the application of democratic faith in the development of the American heritage
- c. Of the forces which have made for world interdependence and the need for world organization
- d. Of the historical and geographic reasons for the behavior of regional and national groups
- e. Of the local community and its problems, and the need for wide participation in community concerns by all citizens
- f. Of the significance in social problems of the mental health and emotional balance of individual human beings

<sup>2</sup> Maurice P. Moffatt and Hazel W. Howell, *Elementary Social Studies Instruction*. Longmans, Green and Co., New York, 1952. p. 12.

<sup>3</sup> *Scope and Sequence of the Social Studies Program*. Wisconsin Cooperative Education Planning Program, Bulletin No. 14. State Department of Public Instruction, Madison, Wis., November 1947. p. 6-7.



## 2. Attitudes

- a. That all human beings regardless of race, national origin, color, or any matter over which they may have no control are entitled to equal rights to life, liberty, and the pursuit of happiness
- b. That we concern ourselves with achieving and improving human welfare and democratic liberties everywhere in the world
- c. That all citizens should participate actively in working toward the solution of community problems for social betterment
- d. That reflective group thinking can serve as an approach toward the solution of social problems

## 3. Skills and/or abilities

- a. The ability to take part in group discussion
- b. The ability to take part in group planning
- c. The ability to think reflectively on social problems
- d. The ability to search out and use valid and adequate sources of information
- e. The ability to evaluate ideas and opinions on controversial problems offered by and through radio, movies, newspapers, periodicals, books, etc.

The student should note that these objectives are listed as understandings, attitudes, skills, and abilities. The best modern thinking in the social studies results in objectives of this definite type rather than in the indefinite and vague objectives that are often listed even today.

### Organization of the social studies

The question of whether to organize the social studies according to the traditional subject divisions, to integrate the various specific subjects into a unified course, or even to integrate the social studies and other areas of knowledge into a core curriculum has received much attention from students in this field. Unified courses are based on the theory that the best way to prepare children to meet the problems they must face in life is to disregard subject divisions and to assemble materials from all sources possible. The core curriculum goes still farther in that it completely ignores traditional subject boundaries.

Believers in the unified course ignore history, geography, and civics as separate subjects and embody material from all of them in a single course. The core curriculum embodies the concept of social education and emphasizes social learning in attaining its goals. There has been

a strong tendency toward an integration of the social studies, especially in the elementary grades, and the tendency has even extended in some degree to the junior- and senior-high-school grades where subject lines have usually been quite clearly drawn.

Although the results of comprehensive investigations show a preponderance of opinion in favor of unification of the social studies in the elementary school and curriculum makers have evolved plans for such an integration, many schools continue to teach subject matter as traditionally organized. Since testing necessarily lags behind the development of the curriculum, there will be a real need for standardized tests and other evaluative devices in the newly organized instruction in this field as classroom practices change.

## 2 OUTCOMES OF SOCIAL STUDIES

### General outcomes of the social studies

It is important that instructional objectives be restated as outcomes in terms of the behaviors developed in pupils. The teacher is better able to measure and evaluate pupil growth in a subject area as complex as the social studies through an understanding of such outcomes than through an understanding of instructional objectives alone.

Wrightstone and Campbell listed six characteristics a pupil should possess if he is to become an effective citizen in a democratic society.\* Their statements, in the form of behavioral outcomes, are that the pupil:

1. Is an individual who is motivated by democratic attitudes and beliefs.
2. Is interested in and sensitive to the problems of the community and of the nation in which he lives.
3. Develops powers of critical and objective thinking.
4. Has suitable work and study skills for acquiring new information and knowledge.
5. Has historical perspectives and concepts and information so that he can make a balanced appraisal of current events, movements, and thought in relation to events which have occurred in the past.
6. Is able to adapt himself to the personal and social conditions which surround and confront him.

\* J. Wayne Wrightstone and Dunk S. Campbell, *Social Studies and the American Way of Life*. Row, Peterson and Co., Evanston, Ill., 1942, p. 25-26.



## Specific outcomes of the social studies

Lists of outcomes such as that given above are still a step removed from the needs of the teacher. Outcomes must be made specific and recognizable to the teacher in terms of definite pupil behavior. The classification outlined below is by Anderson, Forsyth, and Morse.<sup>5</sup>

### A. Acquiring Functional Information.

1. Understanding special vocabulary.
2. Understanding chronological relationships.
3. Understanding maps.
4. Understanding graphs and tables.

### B. Analyzing Social Problems.

1. Knowledge of important concepts, generalizations, and findings.
2. Locating, selecting, organizing, and evaluating information.
3. Drawing conclusions and stating them effectively.
4. Applying social facts, generalizations, and value principles to new problems.

### C. Practicing Desirable Social Relationships.

1. Understanding and developing values consistent with the democratic way of life.
2. Understanding the social implications of specific facts and types of behavior.
3. Applying democratic values . . . in judging the desirability of policies and courses of action.
4. Understanding the importance of social action to further the solution of social problems, and being willing to take such action.

Although the authors of these outcomes outlined methods of measuring and evaluating such behaviors,<sup>6</sup> space in this volume permits only a listing of the outcomes and illustrations of tests and techniques designed to measure some characteristics of these and closely similar types.

## Problems of measuring outcomes

The difficulty of measuring the outcomes of the social studies is great. Thus far, apparently, there has been too little careful analysis

<sup>5</sup> Howard R. Anderson, Elaine Forsyth, and Horace T. Morse, "The Measurement of Understanding in the Social Studies." *The Measurement of Understanding*, Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. University of Chicago Press, Chicago, 1946. p. 72-80. Quoted by permission of the Society.

<sup>6</sup> *Ibid.* p. 80-101.

of the several subjects into the desired knowledges, skills, concepts, understandings, interests, and attitudes to permit exacting curriculum and test construction. Two crucial problems relating to this deficiency are discussed briefly here.

The importance of factual knowledges in the social studies is still a moot question. The modern emphasis upon the development of skills involved in social living and upon the development of work-study skills by the use of which pupils can locate factual knowledges as needed has in part resolved the conflict. Modern schools tend to emphasize carefully selected, functional facts directly useful in the solution of common social problems and to stress concepts, understandings, and abilities to apply facts in problem-solving rather than to teach large numbers of facts indiscriminately. The selection of the facts to teach and which ones to teach as exact and as approximate knowledges continue to be major problems, however. Consequently, the particular facts to test and the degree of knowledge to be expected of pupils continue to be measurement problems.

A further problem exists in the area of the skills necessary for effective social living. Many of them cannot be measured directly in the behavior of the school child because the pertinent behaviors are not often evidenced in the school. They are often revealed in the pupils' out-of-school life and are therefore not subject to direct measurement. They may even be of types for which only the adult behavior of the present school child is the true criterion. Consequently, the problem of measurement and evaluation is the degree to which present school behavior is representative of out-of-school and even adult behavior and the degree to which results from social attitudes tests accurately represent social actions. These questions have not as yet been answered satisfactorily.

### 3 STANDARDIZED SOCIAL STUDIES TESTS

#### Kinds of tests in history, civics, and geography

The selection of the basic facts to be taught and tested is one of the very serious problems of measurement in the social studies. The available body of facts in geography, history, and civics is large and the rapid pace of events today results in constant and great increases in the content of these fields. It is not so much the need for knowledge of the array of facts as it is the determination of those



likely to last long enough in a rapidly changing world to deserve special emphasis in instruction and in testing that complicates the problem. In their efforts to meet the problem of which facts to teach and test, most workers in these fields have made their courses of study and their tests more and more comprehensive, hoping thereby to satisfy the ideas of all concerning the basic items. Too often this has led both teacher and pupils to emphasize mere memorization of extensive catalogs of facts. As a result, these facts are too frequently mastered merely as facts, and not in order that they may give the pupil a better understanding of life and human relationships.

The real deficiency in existing tests in the social studies, however, is not that they are designed primarily to measure the informational aspects of the subject but that other abilities more important to the attainment of the larger objectives of social studies instruction have not been provided for. Unfortunately, when a standardized test is used the particular outcomes it measures tend to be given special attention by both teacher and pupils. As a result, important objectives other than those emphasized in the tests are likely to be neglected.

The majority of the tests available at present in the social science fields are of doubtful value for diagnostic purposes. Three general groups of tests in the social studies may be identified: (1) tests of facts and information, (2) tests of ability to solve social problems, and (3) tests of civic, social, and economic attitudes.

*Factual tests.* Tests of facts and information are by far the most numerous of the tests in the social studies. This is to be expected, for the pupil's knowledge of certain facts or items of information is quite easily discovered. Furthermore, teachers of the social studies have tended to emphasize the acquisition of facts and information to the practical exclusion of other desirable general outcomes of instruction. Factual tests are of limited value for diagnostic purposes. They fail to reveal why pupils do not know the facts if they have not been acquired. The factual tests do not aid the teacher very significantly in discovering the ability of pupils to use facts in their thinking in the social science fields.

*Problem-solving or thought tests.* The development of the ability to utilize facts and basic principles in the attack on a novel social situation is one of the basic outcomes of teaching in the social studies. This type of problem-solving duplicates the steps in the ordinary process of thought. As in arithmetic, problem-solving in the social

studies involves reading the problem to comprehend it, picking out the facts that are pertinent to the problem, choosing a method of solution, and testing the results for accuracy and probability.

It is well recognized that knowledge of the facts necessary for the solution of a problem is no guarantee that the problem will be solved, nor can a problem be solved unless the necessary facts are available. However, availability of facts in this day of widely available library facilities does not depend only upon a knowledge of them by their prospective user. Many of the tests for various types of problem-solving abilities present the necessary facts to the pupils in the test so that the result will depend upon their abilities so to use the facts that they are able to solve the problems.

*Attitudes inventories.* Since actions depend to such a large degree upon attitudes and emotional reactions, the measurement of attitudes resulting from instruction in the social sciences is as greatly needed as are tests of ability to solve problems. As a matter of fact, much attention is now being given in school to the development of the desirable traits of citizenship which are so much needed in later adult life. However, the measurement of such traits by attitudes scales has been largely unrealized, for the attitudes inventories available are in the main much better adapted to the secondary- than to the elementary-school level.

### Standardized tests in social studies subjects

Most of the currently available standardized tests in history, civics and government, and geography were published some years ago. It is largely in the form of a few tests for general social studies and the social studies parts of achievement test batteries that new standardized tests have appeared for this field. This may be the result of the trend toward unification of the social studies in the elementary school.

*History.* Standardized history tests for the elementary and junior-high-school grades are entirely for American history, in order to conform to the course offerings below the high-school level. The major emphasis of most tests is upon factual knowledges, although some of the tests satisfactorily measure some of the more complex and significant results of instruction requiring various applications and interpretations of factual data.



*Civics and government.* Standardized tests in the field of civics and government are limited in number. In general, measurement here is as satisfactory as could be expected under the changing conditions now existing in the social studies. However, there is need for tests that attack important citizenship problems in a more positive and realistic manner than do most of the standardized tests in civics now available.

*Geography.* Many tests are available in geography, but most of these are of the formal factual type. Few of the tests take into account the problem-solving aspects of social studies instruction. The majority of standardized tests in geography attack the subject as a study of places and their characteristics, whereas the modern approach to the study of geography has come largely to be founded upon the manner in which geographical factors influence human beings and the societies they establish.

### Standardized tests in general social studies

Makers of standardized tests have recently been developing tests in the social studies at the junior-high-school and even the elementary-school level to meet the needs of schools that may be offering the unified type of social studies course discussed in a preceding section of this chapter. Tests of this type are not uncommon for the high school, but most of the social studies tests for the elementary- and junior-high-school grades have been for particular courses until the last few years. These tests include material from history, from civics and government, and from geography, but subject-matter lines are broken down. Some of them include content from related sciences as well as from the social studies.

Tests that measure broadly over the social studies must almost of necessity avoid some of the weaknesses of tests in particular subjects because of their lack of concern for divisions within the field. Furthermore, the few tests of this type are relatively new, and consequently have the advantage of being constructed with regard for recent thinking and experimentation with tests. Factual knowledges are less stressed and greater emphasis is placed upon relationships, applications, interpretations, and other reasoned uses of facts than is true on the average of standardized tests for particular courses at the elementary level.

## Interpretive tests in the social studies

The types of tests discussed and illustrated in Chapter 9 as interpretive in nature are more widely available for the senior high school than for the elementary grades and junior high school. They are not dealt with in this and other chapters on measurement and evaluation in subject areas because the authors classify them among the evaluation instruments, tools, and techniques to which Chapter 9 is devoted. Such tests typically cut across the lines of demarcation between subject areas, so that a teacher interested in the measurement of direct outcomes from a course in American history or from a unified course in the social studies would not find them valid for his purposes. However, they measure broad functional outcomes of a type often sought by teachers of social studies and therefore warrant the careful attention of teachers of this subject field.

## Standardized test methods

The practice of presenting illustrative types of objective items to familiarize the student with representative measurement techniques is followed in this chapter. A desirable degree of knowledge on the part of the student concerning specific standardized tests could not be assured in the brief treatment possible here, however. The student can gain such knowledge of particular tests only by examining them critically or even administering them to groups of pupils under standard conditions.

A few sample items representative of test item techniques used by makers of standardized history, civics, geography, and general social studies tests are given in this section. These should serve the double purpose of acquainting students and teachers with standardized testing techniques and of suggesting to them types of test items and exercises they can construct for their own informal objective tests.<sup>7</sup>

*Simple recall items.* The simple recall item is not widely used in standardized social studies tests. Only one sample of this form is presented, for differences among various recall items exist mainly in minor details. The sample is of the basic simple recall form.

<sup>7</sup> See also Anderson, Forsyth, and Morse, *op. cit.* Chapter 5.



Sample A.<sup>8</sup>

- 41 The successor of McKinley to the presidency  
was named ..... (41) \_\_\_\_\_
- 42 The battle cry of the Texan Army was "Re-  
member the ..... (42) \_\_\_\_\_
- 43 The ship in which Henry Hudson first sailed  
up the Hudson River was called the .... (43) \_\_\_\_\_

*Alternate-response items.* The majority of alternate-response items in elementary social studies tests are of the true-false or yes-no variety, although other forms occasionally occur. The illustrations are of a true-false item and a special adaptation of the alternate-response item.

Sample B.<sup>9</sup>

1. Rice grows best in dry soil. **T F 1**
2. Adobe bricks are bricks made of  
clay and dried in the sun. **T F 2**
3. The best way to keep farm soil in  
good condition is to plant the same  
crops each year. **T F 3**

Sample C.<sup>10</sup>

III. Place a cross before the event which came first in each of the following groups:

- 21 ( ) Beginning of Mexican War or  
( ) Annexation of Texas
- 22 ( ) Admission of California as a state or  
( ) Discovery of gold in California

<sup>8</sup> M. H. DeGraff, G. M. Ruch, and H. A. Greene, *Iowa General Information Test in American History*, Grades 7-12. Published by Bureau of Educational Research and Service, University of Iowa, 1927.

<sup>9</sup> Georgia S. Adams and John A. Sexson, *Progressive Tests in Social and Related Sciences, Test 4, Basic Social Processes*, Elementary. Published by California Test Bureau, 1946.

<sup>10</sup> Lena A. Ely and Edith King, *Ely-King Tests in American History*, Junior High School. Published by Southern California School Book Depository, 1927.

*Multiple-choice items.* The multiple-choice item is the most popular for testing purposes in the elementary social studies subjects. The illustrations given below are of the simple item form, of a variation used to test knowledge of vocabulary, and of a variation based on a map.

Sample D.<sup>11</sup>

70. What was the principal work done on a medieval manor?  
 70-1 Defending the castle against attack.  
 70-2 Buying from and selling to caravans.  
 70-3 Manufacturing shoes and cloth.  
 70-4 Farming.  
 70-5 Copying ancient manuscripts. ....70( )
71. The Greek city-states never united, largely because of  
 71-1 geographic barriers.  
 71-2 different languages spoken in different cities.  
 71-3 religious differences.  
 71-4 conflicting forms of government.  
 71-5 opposition from Persia. ....71( )

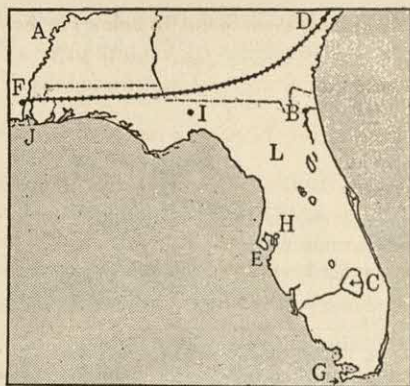
Sample E.<sup>12</sup>

46. substitute    <sup>1</sup> food    <sup>2</sup> clothing  
                   <sup>3</sup> substance    <sup>4</sup> replace    \_\_\_\_\_46
47. erosion    <sup>1</sup> proposition    <sup>2</sup> alliance  
                   <sup>3</sup> disintegration    <sup>4</sup> concession    \_\_\_\_\_47
48. nullify    <sup>1</sup> pacify    <sup>2</sup> return  
                   <sup>3</sup> cancel    <sup>4</sup> repeat    \_\_\_\_\_48
49. humidity    <sup>1</sup> cupidity    <sup>2</sup> disposition  
                   <sup>3</sup> moisture    <sup>4</sup> secretion    \_\_\_\_\_49
50. league    <sup>1</sup> disease    <sup>2</sup> alliance  
                   <sup>3</sup> departure    <sup>4</sup> pattern    \_\_\_\_\_50

<sup>11</sup> Harry D. Berg and Elaine Forsyth, *Cooperative Social Studies Test for Grades 7, 8, and 9*. Published by Cooperative Test Service, 1947.

<sup>12</sup> Ernest W. Tiegs and Willis W. Clark, *California Reading Test, Reading Vocabulary in Social Science*, Advanced. Published by California Test Bureau, 1950.



Sample F.<sup>13</sup>

16. Which one of the following letters shows where there is a seaport indicated on the map?  
 F I D B ..... ( ) 16
17. Which letter shows where there is a bay?  
 L E C A ..... ( ) 17
18. Which letter shows where there is a cape?  
 F H L G ..... ( ) 18

*Matching exercises.* Matching exercises appear to be second to multiple-choice item forms in popularity for the testing of achievement in elementary social studies courses. Not only are balanced matching exercises widely used, but exercises based on the multiple use of one category of items and on graphs or maps are also common. Illustrations are given below for two of these types.

Sample G (Exercise shown only in part).<sup>14</sup>

- |                   |   |
|-------------------|---|
| Champlain (1)     | 36. Pioneered in Kentucky ..... ( )         |
| James Wolfe (2)   | 37. Captured Fort Ticonderoga ..... ( )     |
| Ethan Allen (3)   | 38. Helped to settle Jamestown ..... ( )    |
| John Winthrop (4) | 39. Famous missionary to the Indians... ( ) |

<sup>13</sup> Richard D. Allen and others, *Metropolitan Achievement Tests, Test 8, Social Studies: Geography*, Intermediate. Published by World Book Co., 1946.

<sup>14</sup> E. C. Denny and M. J. Nelson, *Denny-Nelson American History Test*, Grades 7 and 8. Published by World Book Co., 1928.

Sample H (Exercise shown only in part).<sup>15</sup>

DIRECTIONS. After each event in the list below put the number —

- 1 if it happened before the *Settling of Jamestown* in 1607.
- 2 if it happened between the *Settling of Jamestown* and the *Adoption of the Constitution* in 1787.
- 3 if it happened between the *Adoption of the Constitution* and the *Civil War* in 1861-65.
- 4 if it happened between the *Civil War* and the *Spanish-American War* in 1898.
- 5 if it happened since the *Spanish-American War*.

For example, you should write the number 1 after "Columbus discovered America," because it happened before the *Settling of Jamestown*.

SAMPLE. Columbus discovered America ..... ( )

- |   |     |    |
|---|-----|----|
| 42. The Stamp Act.....  | ( ) | 42 |
| 43. The South Pole was discovered.....                        | ( ) | 43 |
| 44. Eli Whitney's cotton gin was perfected.....               | ( ) | 44 |
| 45. The first trip around the world was completed.....        | ( ) | 45 |
| 46. Daniel Boone guided settlers across the Appalachians..... | ( ) | 46 |

## 4 CLASSROOM TESTING AND EVALUATING IN SOCIAL STUDIES

### Informal objective tests in the social studies

More attention has been given to informal objective testing methods for the social studies of the high-school level than of the elementary-school level in the educational literature, except possibly for geography. This suggests that few new techniques or modifications of old techniques have been devised for the social studies below the high school. Such a situation is not surprising, in view of the fact that the small number of new standardized tests in this field are included in achievement test batteries or are for the general social studies.

Two means of evaluating instructional outcomes of the social studies informally are open to the teacher: (1) the construction of informal objective tests, and (2) the use of other evaluative procedures. Illustrations and discussions of item types in the preceding section of this chapter should aid the teacher in constructing objective classroom tests. The program of evaluation quoted below deals largely with devices of a non-test nature.

<sup>15</sup> Richard D. Allen and others, *Metropolitan Achievement Tests, Test 7, Social Studies: History and Civics*, Advanced. Published by World Book Co., 1946.



## An evaluation program

A comprehensive program for the evaluation of the instructional outcomes of the social studies is given below for its value in suggesting a variety of suitable measurement techniques to supplement paper-and-pencil tests. Wesley and Adams pointed out that although most of the suggested techniques are objective, materials for all of them are not in existence. A challenge is thereby presented to the teacher to devise his own evaluation instruments in such cases.

### A PROGRAM OF EVALUATION <sup>16</sup>

<i>Outcomes</i>	<i>Evaluative Technique</i>
1. Concepts	Ability to use the word, to choose the correct synonym, to define it, to match it with an example or definition; presence of the word in the pupil's written and spoken vocabulary; number of denotations and connotations of a particular word which the pupil knows.
2. Interests	Record of books chosen; record of the parts of a newspaper read without guidance; nature of magazine articles read; nature of conversations; choice of games, projects, and problems; shows attended; hobbies; radio programs; use of leisure time.
3. Cooperation	Records of instance of cooperation; test of attitude toward cooperation; number of pupil organizations to which he belongs; the quality and extent of his contribution; instances of his lending balls, pencils, etc.; willingness to take turns; number of friends.
4. Toleration	Attitude toward shy and backward pupils; respect for other peoples' opinions; naturalness in presence of pupils who are "different"; instances of kindness.
5. Locating Materials	Directness with which a pupil finds needed books and type selected; tests of familiarity with yearbooks, atlases, encyclopedias, etc.; time test in use of index and dictionary; tests of discrimination in choice of sources.

<sup>16</sup> Edgar B. Wesley and Mary A. Adams, *Teaching Social Studies in Elementary Schools*. D. C. Heath and Co., Boston, 1946. p. 349-50.

- |                            |  |
|----------------------------|--|
| 6. Studying<br>Materials   | Ability to select main ideas; recognition of symbols, and abbreviations; kinds of notes and notebooks; speed and legibility of writing; method of using globes and maps; interpretation of pictures, graphs, and cartoons; ability to outline and summarize.   |
| 7. Appraising<br>Materials | Ability to distinguish between sources and secondary accounts; degrees of reliability; degrees of probability; discrimination among writers; lists of preferred books, shows, meetings; ability to sense inconsistencies, to distinguish facts and opinions; degree of difficulty in proving different kinds of statements; recognition of tentative nature of conclusions in social studies; awareness of conflicting statements; selection of relevant data. |
| 8. Utilizing<br>Materials  | Ability to make proper deduction from a generalization; ability to make logical inference, state a conclusion, draw a conclusion; ability to write a report; ability to present a report orally; ability to select a suitable map, graph, or picture for a particular purpose; ability to sort ideas into proper categories.   |

## 5 CORRECTIVE WORK IN SOCIAL STUDIES

### Diagnosis and remedy in the social studies

Diagnosis in the social studies is difficult because (1) the knowledges, skills, and understandings that pupils should acquire are not too clearly identified, and (2) if the facts to be learned were known accurately it would still be impossible to determine whether the pupil functioned in his social relationships in a desirable manner because of his possession of the informational elements revealed by a test. Diagnosis and remedy are often needed in those skills that are basic to successful work in the social studies. Instruction in these subjects requires much reading of the work-study type. Therefore, pupils, in order to achieve at acceptable levels, must possess many of the following work-study reading skills:

1. Knowledge of technical vocabularies employed in the social studies.
2. Reading comprehension adequate for interpretation of social science content.



3. Ability to locate material readily—use of the index, library files, table of contents, maps, charts, etc.
4. Ability to outline.
5. Ability to summarize.

These skills are discussed in Chapter 15, along with other ways and means for corrective work in these important acquisitive skills. They are not, therefore, taken up here.

## Topics for Discussion

1. Define the field of the social studies in such a way as to clarify the objectives adequately for testing purposes.
2. Discuss the pros and cons of a unified social studies curriculum as contrasted with the traditional organization of the content by subjects.
3. State four of the outcomes of instruction in the social studies.
4. What are the three main types of social studies tests as specified in this chapter?
5. In your judgment what is the relation of factual knowledges to ability in problem-solving in the social studies?
6. What are the chief weaknesses in problem-solving tests and attitudes scales?
7. Discuss the use of various objective test item forms in standardized social studies tests.
8. Comment on some of the evaluative techniques of a non-test nature suggested for use in the social studies.

## Selected References

- ANDERSON, HOWARD R. "Classroom Evaluation of the Awareness of Propaganda." *Education against Propaganda*. Seventh Yearbook of the National Council for the Social Studies. Washington, D. C.: The Council, 1937. p. 171-82.
- ANDERSON, HOWARD R. "Examinations in the Social Studies." *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Co., 1936. Chapter 4.
- ANDERSON, HOWARD R. "Testing in the Social Studies." *Education*, 58:545-49; May 1938.
- ANDERSON, HOWARD R., AND LINDQUIST, E. F. *The Improvement of Objective Testing in History*. Second Yearbook of the National Council for the Social Studies. Philadelphia: McKinley Publishing Co., 1932. p. 97-117.

- ANDERSON, HOWARD R., FORSYTH, ELAINE, AND MORSE, HORACE T. "The Measurement of Understanding in the Social Studies." *The Measurement of Understanding*. Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1946. Chapter 5.
- BARNETT, SIDNEY W. "Testing for Objectives in the Social Studies." *High Points*, 29:56-68; December 1947.
- BROOM, M. E. *Educational Measurements in the Elementary School*. New York: McGraw-Hill Book Co., Inc., 1939. Chapter 9.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 657-73.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p. 404-28.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938. p. 67-72, 82-83, 88-92, 149-52.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. p. 604-26.
- CAIN, MAUD. "A Study of Thirteen Standard Geography Tests." *Journal of Geography*, 34:252-56; September 1935.
- CARR, EDWIN R., AND OTHERS. "Social Studies." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 1213-38.
- HAMALAINEN, ARTHUR E. "Evaluation in the Social Studies." *Social Studies*, 28:250-52; October 1937.
- HERRICK, JOHN H. "The Evaluation of Certain Aspects of Thinking in the Social Studies." *Educational Method*, 15:422-26; May 1936.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. Chapter 7.
- KELLEY, TRUMAN L., AND KREY, AUGUST C. *Tests and Measurements in the Social Sciences*. Report of the Commission on the Social Studies, American Historical Association, Part IV. New York: Charles Scribner's Sons, 1934.
- KOHN, CLYDE F. "Geography." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 501-3.
- MICHAELIS, JOHN U. *Social Studies for Children in a Democracy*. New York: Prentice-Hall, Inc., 1950. Chapter 15.
- MOFFATT, MAURICE P. *Social Studies Instruction*. New York: Prentice-Hall, Inc., 1950. Chapter 14.
- MORSE, HORACE T., editor. "Evaluation and Tests in American History." *The Study and Teaching of American History*. Seventeenth Yearbook of the National Council for the Social Studies. Washington, D. C.: The Council, 1947. Section 6.



- ORATA, PEDRO T. "Evaluation in the Field of Social Science." *Educational Method*, 16:121-37; December 1936.
- TYLER, RALPH W. "Improving Test Materials in the Social Studies." *Educational Research Bulletin*, 11:373-79; November 9, 1932.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapters 14-15.
- WEITZMAN, ELLIS, AND MCNAMARA, WALTER J. *Constructing Classroom Examinations*. Chicago: Science Research Associates, 1949. p. 63-69.
- WESLEY, EDGAR B. "Diagnosis in the Social Studies." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1935. Chapter 15.
- WESLEY, EDGAR B., AND ADAMS, MARY A. *Teaching Social Studies in the Elementary School*. Boston: D. C. Heath and Co., 1946. Part VII.
- WRIGHTSTONE, J. WAYNE. "Measuring Some Major Objectives of the Social Studies." *School Review*, 43:771-79; December 1935.
- WRIGHTSTONE, J. WAYNE, AND CAMPBELL, DOAK S. *Social Studies and the American Way of Life*. Evanston, Ill.: Row, Peterson and Co., 1942. Part III.

## ***Measuring and Evaluating in Elementary-School Mathematics***

THE FOLLOWING important points involved in the measurement and evaluation of skills in elementary-school mathematics are summarized in this chapter:

- A. Modern ideas concerning the nature of elementary-school mathematics.
- B. Importance of meanings and understandings in the development of computational skills.
- C. Desirable outcomes of instruction in arithmetic.
- D. Measurement of general achievement in arithmetic.
- E. Diagnostic testing in arithmetic.
- F. Remedial instruction in arithmetic.

The content of elementary-school mathematics, which to most individuals means arithmetic, is not affected seriously by social and scientific changes, as is the case in certain other subjects. Mathematics is based upon a fixed system of numbers that bear definite relations to each other. In the decimal system, two *times* three always equals six and two *plus* three always equals five. New discoveries and recent research are not likely to change these basic facts. However, definite as the facts of arithmetic may be, there are many evidences of changes in current beliefs about the importance of the subject, the best methods of mastering it, and the most effective ways of relating it to other important social objectives of education.

Within the past generation the emphasis on certain of the previ-



ously acceptable instructional objectives of elementary-school mathematics seems to have shifted significantly. While the earlier purpose of developing arithmetical skill to a point of computational efficiency still remains as one important objective, the modern elementary school is primarily concerned with securing real competence in the use of computational skills, not as an end in itself but as a means of understanding the quantitative aspects of the world about us. This broadening and shifting of ideas concerning the functions of arithmetic in the elementary school is nowhere more clearly indicated than by the current tendency to substitute the term *elementary-school mathematics* for the older term in modern literature on the curriculum.

Computational skills as factors in social efficiency may still be the desired but not too-often-realized goals of instruction in arithmetic. Critical students of elementary-school mathematics are seriously concerned with the development of understandings and competence in quantitative thinking often involving little or no computational exactness. These practical objectives have in turn produced definite changes in the methods and materials of instruction in the subject. The basic facts themselves upon which computational skill rests may be mostly unchanged, but many of the procedures and devices used to develop meanings and understandings on the part of the individual are new and in many ways quite different from those in use a decade ago.

## 1 COURSE CONTENT AND ORGANIZATION IN ARITHMETIC

Three principal methods of approach to the question of what should be taught in arithmetic have been used: (1) social usage, (2) child usage, and (3) the psychological approach. The numerous studies of social usage of arithmetic have resulted in many modifications of the curriculum, chiefly through the elimination of such non-functional skills as cube root and cases in percentage, in order to adapt instruction more accurately to the needs individuals encounter for arithmetic skills and abilities. The child-usage approach is similar, although it is most widely useful for the first two grades during which informal number activities rather than formal instruction constitute the work in arithmetic. Both of these procedures are based on studies of social utility, which determine the types of arithmetic skills and abilities people actually have occasion to use in real-life

situations. The third basis for determining the content and organization of the arithmetic curriculum supplements rather than conflicts with the first two. Based on the unit skills important in the subject, it might be called the psychological approach.

## Organization of arithmetic instruction

Of these three methods, the first was for a time responsible for a better selection of the types of arithmetic skills taught in the schools and the third has been chiefly responsible for the organization of content for teaching and remedial purposes. Child-usage studies have contributed mainly to instruction in the primary grades and to a certain extent in the activity school.

Application of the social utility theory eventually resulted in such a great reduction of content in arithmetic and in the time devoted to its teaching that specialists began to question the advisability of any further continuance of the reductionist trend and to recommend enrichment of the curriculum by the addition of new content. Knight very early expressed this attitude by pointing out that an improved curriculum cannot result from a process of subtraction only, but that the determination of what must be added to the course of study is of comparable importance.<sup>1</sup>

The drill method of teaching, which assumed that arithmetic facts and skills are largely unrelated and should consequently be taught in isolation, also came to be questioned seriously. Overman, in summarizing various studies of the degree to which arithmetic skills are transferred by pupils to numerical situations not previously encountered, concluded that instruction which stresses general rules, relationships, and methods of procedure is more effective than that which stresses isolated facts and skills.<sup>2</sup>

Brownell summarized the meaning theory of arithmetic instruction, which has emerged during the past two decades, by stating that, within the meaning theory there is absolutely no place for the view of arithmetic as a heterogeneous mass of unrelated elements to be trained

<sup>1</sup> Frederick B. Knight, "Some Considerations of Method." *Report of the Society's Committee on Arithmetic*. Twenty-Ninth Yearbook of the National Society for the Study of Education. Public School Publishing Co., Bloomington, Ill., 1930. p. 149.

<sup>2</sup> James R. Overman, "The Problem of Transfer in Arithmetic." *The Teaching of Arithmetic*, Tenth Yearbook of the National Council of Teachers of Mathematics. Bureau of Publications, Teachers College, Columbia University, New York, 1935. p. 179-80.



through repetition. The meaning theory conceives of arithmetic as a closely knit system of understandable ideas, principles, and processes. According to this theory, the test of learning is not mere mechanical facility in 'figuring.' The true test is an intelligent grasp upon number relations and the ability to deal with arithmetical situations with proper comprehension of their mathematical as well as their practical significance.<sup>3</sup>

The meaning approach does not abandon drill as a teaching device nor the results from psychological analyses of basic skills in the organization of instruction, but it attempts so to organize instruction that quantitative relationships become meaningful to the child. For example, Brownell pointed out that common fractions, decimal fractions, and percentage, commonly taught as different mathematical forms, are actually but three different ways of expressing the same ideas, and should be taught in that manner.<sup>4</sup> Spitzer, in a much more recent summary of theoretical and experimental considerations affecting learning in arithmetic, stated that "The most effective learning procedures emphasize meaning and understanding."<sup>5</sup> Later in this same discussion he pointed out that "It is advantageous to the learner to see the reasons for the use or the application of the arithmetic that is being studied. . . . Drill on fundamental phases of arithmetic is essential if the needs of children are to be met . . . To be effective, drill should follow understanding."<sup>6</sup>

### Desirable outcomes of instruction in arithmetic

In a discussion of the place of arithmetic and its relationship to the total elementary-school curriculum, Horn<sup>7</sup> concluded that the following statement of desirable arithmetic outcomes as given by Brownell is entirely in line with modern instructional trends:

<sup>3</sup> William A. Brownell, "Psychological Considerations in the Learning and the Teaching of Arithmetic." *The Teaching of Arithmetic*, Tenth Yearbook of the National Council of Teachers of Mathematics. Bureau of Publications, Teachers College, Columbia University, New York, 1935. p. 19.

<sup>4</sup> *Ibid.* p. 26.

<sup>5</sup> Herbert F. Spitzer, "Learning and Teaching Arithmetic." *The Teaching of Arithmetic*, Fiftieth Yearbook of the National Society for the Study of Education, Part II. University of Chicago Press, Chicago, 1951. p. 141.

<sup>6</sup> *Ibid.* p. 141-42.

<sup>7</sup> Ernest Horn, "Arithmetic in the Elementary School Curriculum." *The Teaching of Arithmetic*, Fiftieth Yearbook of the National Society for the Study of Education, Part II. University of Chicago Press, Chicago, 1951. p. 6.

OUTCOMES OF INSTRUCTION IN ARITHMETIC IN THE  
ELEMENTARY SCHOOL

## 1. Computational skill:

Facility and accuracy in operations with whole numbers, common fractions, decimals, and per cents. (This group of outcomes is here separated from the second and third groups which follow because it *can* be isolated for measurement. In this separation much is lost, for computation without understanding *when* as well as *how* to compute is a rather empty skill. Actually computation is important only as it contributes to social ends.)

## 2. Mathematical understandings:

- a) Meaningful conceptions of quantity, of the number system, of whole numbers, of common fractions, of decimals, of per cents, of measures, etc.
- b) A meaningful vocabulary of the useful technical terms of arithmetic which designate quantitative ideas and the relationships between them.
- c) Grasp of important arithmetical generalizations.
- d) Understanding of the meanings and mathematical functions of the fundamental operations.
- e) Understanding of the meanings of measures and of measurement as a process.
- f) Understanding of important arithmetical relationships, such as those which function in reasonably sound estimations and approximations, in accurate checking, and in ingenious and resourceful solutions.
- g) Some understanding of the rational principles which govern number relations and computational procedures.

## 3. Sensitiveness to number in social situations and the habit of using number effectively in such situations:

- a) Vocabulary of selected quantitative terms of common usage (such as kilowatt hour, miles per hour, decrease and increase, and terms important in insurance, investments, business practices, etc.).
- b) Knowledge of selected business practices and other economic applications of number.
- c) Ability to use and interpret graphs, simple statistics, and tabular presentations of quantitative data (as in study in school and in practical activities outside of school).
- d) Awareness of the usefulness of quantity and number in dealing with many aspects of life. Here belongs some understanding of



social institutions in which the quantitative aspect is prominent, as well as some understanding of the important contribution of number in their evolution.

- e) Tendency to sense the quantitative as part of normal experience, including vicarious experience, as in reading, in observation, and in projected activity and imaginative thinking.
- f) Ability to make (and the habit of making) sound judgments with respect to practical, quantitative problems.
- g) Disposition to extend one's sensitiveness to the quantitative as this occurs socially and to improve and extend one's ability to deal effectively with the quantitative when so encountered or discovered.<sup>8</sup>

## Basic arithmetic skills

Arithmetic is one of the more definite tool subjects, and much of its content is suitably organized for teaching purposes. For years it has been recognized that success in addition depends on a mastery of the basic addition facts. The same may be said of each of the four fundamental processes with whole numbers. Teachers now recognize, however, that success in such work as long division is dependent on a great many more skills than are involved in the mastery of the basic division facts. Long division calls for the accurate use of skills in addition, multiplication, and subtraction, not to mention the skills that are usually recognized as belonging definitely to division. Multiplication itself may involve the basic multiplication facts, the addition and multiplication involved in carrying in multiplication, and addition itself. A partial catalog of arithmetical skills selected for teaching, testing, and remedial purposes is presented here to illustrate the extent to which such an analysis may be carried as well as to furnish a broad basis upon which to build diagnostic and remedial material in this field.

### BASIC ARITHMETIC SKILLS

#### I. Fundamental Processes with Whole Numbers

- 1. Basic Addition Facts
- 2. Basic Subtraction Facts
- 3. Basic Multiplication Facts

<sup>8</sup> William A. Brownell, "The Evaluation of Learning in Arithmetic." *Arithmetic in General Education*, Sixteenth Yearbook of the National Council of Teachers of Mathematics. Bureau of Publications, Teachers College, Columbia University, New York, 1941. p. 231-32.

4. Basic Short Division Facts
5. Higher Decade Addition
6. Column Addition
7. Carrying in Column Addition
8. Harder Subtraction
9. Borrowing or Carrying in Subtraction
10. Addition Used in Harder Multiplication
11. Carrying in Addition Used in Harder Multiplication
12. Complete Process of Multiplication
13. Short Division Involving Carrying
14. Multiplication, Addition, and Subtraction Used in Long Division
15. Complete Process of Long Division

## II. Fundamental Processes with Fractions and Whole Numbers

1. Changing Fractions to Equivalent Forms
2. Finding Common Denominators
3. Reducing Fractions
4. Addition of Fractions and Mixed Numbers
5. Expressing Mixed Numbers as Improper Fractions
6. Fundamentals of Subtraction of Fractions
7. Reduction of Mixed Numbers
8. Cancellation in the Multiplication of Fractions
9. Multiplication of Fractions
10. Cancellation in Division of Fractions
11. Changing from Multiplication to Division Form
12. Fundamentals of Division of Fractions

## III. Fundamental Processes with Decimals

1. Notation of Decimals
2. Changing Fractions and Mixed Numbers to Decimal Form
3. Changing Decimals to Fractions and Mixed Numbers
4. Fundamentals of Addition of Decimals
5. Fundamentals of Subtraction of Decimals
6. Pointing off in Multiplication of Decimals
7. Dividing Decimals by Pointing off
8. Location of Decimal Points in Division
9. Changing Remainders to Decimal Form
10. Fundamentals of Division of Decimals

## IV. Fundamental Processes with Denominate Numbers

1. Reducing in Denominate Numbers
2. Borrowing in Denominate Numbers
3. Addition of Denominate Numbers



4. Subtraction of Denominate Numbers
5. Multiplication of Denominate Numbers
6. Division of Denominate Numbers

#### V. Mensuration

1. Mensuration of Plane Surfaces
2. Mensuration of Solids
3. Finding Areas and Volumes
4. Formulas Used in Mensuration

#### VI. Percentage

1. Fractional and Per Cent Relations
2. Decimal and Per Cent Relations
3. Expressing Areas in Per Cents
4. Fundamentals of Work in Percentage

#### VII. Interest

1. Business Forms
2. Budgets
3. Computation of Interest
4. Computation of Discount
5. Use of Interest Tables

#### VIII. Problem Solving

1. Comprehension of Problem
2. Knowledge of What Is Given
3. Knowledge of What Is Called for
4. Probable Answer
5. Knowledge of Proper Processes and Proper Order of Processes
6. Recognition of the Correct Solution

## 2 MEASUREMENT OF GENERAL ACHIEVEMENT IN ARITHMETIC

A comprehensive understanding of standardized tests and their nature and use is best attained by an examination of sample tests and, if possible, such accompanying materials as manuals of directions, scoring keys, and pupil record forms, or, preferably, the actual use of one or more tests in the classroom. Therefore, the authors have chosen to present only a few sample items from various tests to illustrate the application of different objective testing methods. To conserve space, directions to the pupils are not given for the sample items except in instances of unusually complex item forms.

It is believed that the presentation of sample items with brief comments will serve two valuable purposes: (1) familiarize the student with information concerning standardized testing methods and major item techniques in arithmetic, and (2) furnish him with suggestions concerning some of the methods he may very well apply in constructing tests for use with his own classes.

### Standardized testing in computational skills

Computational skills are most often tested by an item type of simple recall form, although multiple-choice items are sometimes used. Such item types can be used with any combination of the four fundamental operations—addition, subtraction, multiplication, and division—and the four types of numbers—whole numbers, mixed numbers, fractions, and decimals. Some tests classify all items of a type together, while others use the “omnibus” arrangement of mixed order for the various operations and types of numbers.

*Simple recall items.* Although these items are of simple recall form, it is by means of performing certain calculations rather than as recall that a pupil obtains the answers. Directions are usually given to the pupil concerning the form of answer desired, e.g., mixed numbers reduced to whole numbers and fractions, fractions reduced to lowest terms. Definite rules are also usually provided in order to objectify the scoring of a type of performance that is often viewed by different teachers according to very different standards. Credit is ordinarily given only for answers that are entirely correct.

Sample A.

1. Subtract	2. Subtract	3. Multiply	4. Divide	5. Add	1. _____
58	456	105	$56 \div 7 =$	31	2. _____
-37	-107	6		17	3. _____
_____	_____	_____		24	4. _____
				_____	5. _____

*Multiple-choice items.* Multiple-choice items require the pupils to perform the calculations in order to determine which is the correct answer, although there is usually no requirement that the pupil put down the work by which he obtained the answer. Some pupils might obtain the answers by mental computation and others by putting down only a skeleton of their computations.



Sample B.<sup>9</sup>

41 Add	407	42 Subtract	43 Multiply
	819	6002	265
	965	<u>3918</u>	<u>300</u>
	<u>113</u>		
41 <input type="checkbox"/> 2204	<input type="checkbox"/> 2305	<input type="checkbox"/> 2314	<input type="checkbox"/> N
42 <input type="checkbox"/> 2084	<input type="checkbox"/> 2114	<input type="checkbox"/> 3084	<input type="checkbox"/> N
43 <input type="checkbox"/> 7950	<input type="checkbox"/> 78,500	<input type="checkbox"/> 79,500	<input type="checkbox"/> N

### Standardized testing in problem-solving

Standardized tests in problem-solving are most frequently set up either in simple recall or in multiple-choice form. The five examples given below are sufficient to illustrate the testing method because of the similarity of problem-solving items in different tests.

*Simple recall items.* Simple recall items in this situation require solutions of the problems, rather than recall in the usual sense of that word, in obtaining the answers. Scoring of responses is practically always on an all-or-none basis, for no credit is given unless the answer is correct. Only one illustration of this item type is shown.

Sample C.<sup>10</sup>

1. I bought an apple for 4 cents, a bowl of soup for 8 cents, and a cookie for 2 cents. All of the food cost how many cents? .....
2. John has 6 cents and wants to buy a ball that costs 15 cents. How many more cents does he need to buy the ball? .....

*Multiple-choice items.* The multiple-choice item in problem-solving also usually requires the solution of the problem in order to determine which of the alternative answers is the correct one. However, some items require only an indication of the information necessary in a problem situation.

<sup>9</sup> E. F. Lindquist and others, *Iowa Every-Pupil Tests of Basic Skills, Test D, Basic Arithmetic Skills*, Advanced, Form Q. Published by State University of Iowa, 1946.

<sup>10</sup> Gertrude H. Hildreth, *Arithmetic Achievement Tests*, Grades 2 to 6. Published by Bureau of Publications, Teachers College, Columbia University, 1935.

Sample D.<sup>11</sup>

Jean sold boxes of home-made fudge at 50¢ a box. Each box cost 30¢ to make and contained two dozen pieces.

74. What was her profit on 8 boxes?
75. What was the cost of making each piece?
76. Her profit was what per cent of the selling price?

74	<input type="checkbox"/> \$1.10	<input type="checkbox"/> \$1.60	<input type="checkbox"/> \$4.00	<input type="checkbox"/> N
75	<input type="checkbox"/> 1¼¢	<input type="checkbox"/> 2.1¢	<input type="checkbox"/> 2½¢	<input type="checkbox"/> N
76	<input type="checkbox"/> 40%	<input type="checkbox"/> ¼%	<input type="checkbox"/> 140%	<input type="checkbox"/> N

### 3 DIAGNOSTIC TESTING IN ARITHMETIC SKILLS

Tests as such are incapable of improving instruction directly. Existing conditions are merely revealed by them, and it is worthy of note that these conditions are revealed only within the limits of the validity and the reliability of the particular tests used. The importance of using tests that are themselves based upon a sufficiently detailed analysis of the skills required for successful achievement in the field to permit the application of definite remedial procedures can hardly be overemphasized. Remedial teaching is the result of deliberate instructional effort on the part of the teacher after the particular points of weakness of the pupils have been revealed. The accuracy with which these needs are revealed by the device used is the best measure of its value to the classroom teacher.

#### Scope of diagnostic testing in arithmetic

It is not by chance that diagnostic tests have been developed in subject fields in which the aims are clean cut and the basic skills conditioning achievement have been analyzed carefully. Nor is it by chance that the blanket purposes of certain other subject fields, as expressed in courses of study and textbooks, have left the teacher groping vaguely for tangible goals and effective instructional methods. The order of development is clear: first, there must be a specific statement of aims lying back of the subject; second, a de-

<sup>11</sup> Lindquist, *op. cit.*



tailed analysis must be made of the basic skills upon which ultimate achievement depends; and third, material designed to give mastery of these skills must be prepared.

Some progress has been made in the diagnosis of pupil defects in the field of arithmetic. This is possible because the aims of arithmetic are quite clearly stated, which in turn permits a rather detailed analysis of the underlying skills. As soon as it became known, for example, that the ability to do a certain type of column addition depends on the pupil's knowledge of certain higher-decade addition facts, it was possible not only to locate difficulties in teaching this material as such but also to furnish the teacher with specific aids in teaching it. The reason why similar material is not available in geography, history, and science is that the aims of instruction in these fields have not yet become sufficiently crystallized to permit the type of analysis to which arithmetic has been subjected.<sup>12</sup>

### Diagnostic tests in basic arithmetic skills

There are three currently useful diagnostic tests in arithmetic, each representing a rather specific point of view in diagnosis. None of these tests is of recent copyright, since not a great deal of work has been done during recent years on diagnostic testing in arithmetic. The *Buswell-John Diagnostic Chart for Fundamental Processes in Arithmetic* is designed for individual diagnostic work. It consists of a diagnostic chart and a test sheet on which the pupil does his work aloud in the presence of the teacher. On the diagnostic chart, which is for the teacher's use, are listed the most frequent faulty habits of work and causes of error in the particular arithmetic process under diagnosis. The pupil is given the work sheet and instructed to work on each of the exercises. In this way the teacher is able to discover the pupil's method of work and check the major causes of his difficulty.

<sup>12</sup> It should probably be pointed out here that very likely there are certain subject fields in which this type of crystallization of aims will not and should not take place. This is undoubtedly true of certain subjects, such as social studies and natural and physical sciences, in which changes in content appear rapidly and in which there are certain points on which general agreement cannot be expected. Here the diagnosis will remain for some time in general terms, such as the ability to read or the ability to do the mathematics involved in certain science fields. Here also remedy will be largely in terms of bringing about a more adequate mastery of certain definite materials.

The *Brueckner Diagnostic Tests*, which cover whole numbers, fractions, decimals, and percentage, are really inventory exercises which make it possible for a sufficiently critical and analytical teacher to check through the pupil's work and discover the apparent causes of difficulty. The diagnosis thus becomes a matter of working out an individual analytical record for each child.

The *Compass Diagnostic Tests* represent the third of the approaches to diagnosis. This series consists of twenty tests covering the fundamental processes with whole numbers, fractions, decimals, percentage, arithmetical definitions and concepts, business forms, mensuration, and problem analysis and problem-solving. The tests are essentially analytical in structure, the total process in each case being broken down one step at a time as a basis for the identification of the causes of weakness. These tests are designed for group measurement and diagnosis. It is probable that no diagnostic test in any field is capable of indicating precisely *why* a skill breaks down, but there is not much greater certainty that the teacher's interpretation of why the pupils encountered difficulty will be more exact. The list of skills enumerated in the outline on pages 486 to 488 gives a very definite idea of the wide range of ability that must be covered by such tests.

## 4 TESTING PROBLEM-SOLVING ABILITY

### Meaning of problem-solving

Problem-solving in arithmetic is a complicated skill which is almost certainly highly related to general intelligence. Naturally, an attempt to analyze and to identify the underlying skills meets with considerable difficulty. Thus far five fundamental steps in problem-solving, closely paralleling the steps in the thinking process outlined by Dewey,<sup>13</sup> have been identified. These often-quoted steps afford practically the only workable basis for an attack upon problem-solving difficulties.

The first step in the solution of verbal problems demands a complete understanding of the elements and processes that are involved or implied. This is *comprehension*. This in itself involves many fac-

<sup>13</sup> John Dewey, *How We Think*. D. C. Heath and Co., Boston, 1910.



tors, such as rate of reading, vocabulary difficulties, reading of numerals, and problem organization, as well as complexity in terms of the number and order of the arithmetical processes involved. Underlying all of these is, of course, the ability of the child to hold the various facts and conditions in his mind long enough to analyze and organize them. This process of *analysis and organization* constitutes a second important step. The unnecessary facts or implications are discarded and only the significant data are retained. The third step in practice is actually a part of the second, for the *recognition of the process* involved is really a part of analysis. From this the worker moves straight to the fourth step, *solution*, where he applies to a specific situation his knowledge of the fundamental tools of number. In his earlier practice he has learned how to perform certain simple arithmetical computations. Now he learns when to apply them. The next and final step in the process is *verification*, which may be either a rough checking by the estimation of the probable answer to the problem or an actual recalculating and rechecking of the processes involved.

### Diagnostic tests of problem-solving ability

Naturally, many children fail to proceed in the solution of problems in the orderly fashion indicated in this discussion, although it would ordinarily be economical for them to do so. Oftentimes imperfect work (though finally successful) means using unnecessary steps and spending useless energy in doing essential steps in an ineffective order.

So far in the development of diagnostic instruments for the identification of the skills involved in problem-solving only a rough sampling of these skills has been approximated, and these are found in tests of such ancient copyright dates that they are not reviewed here.

## 5 REMEDIAL INSTRUCTION IN ARITHMETIC

### Remedial materials in arithmetic fundamentals

Pupils do not fail in arithmetic in a vague, general sense, nor do they need remedial work of a vague and general type. Pupils' errors and failures are specific. The more exactly they can be located, the

more promptly they can be removed. Diagnostic tests based upon a satisfactory analysis of the skills that are essential to pupil mastery are for the purpose of locating these specific breakdowns.

Remedial exercises incorporating most of the desirable characteristics of such material can be developed by the classroom teacher or can often be secured in commercial form. The preparation and use of such material to supplement available instructional devices should serve to increase the efficiency of teaching greatly. It should be remembered that the most effective use of remedial material will follow the careful diagnosis of individual pupil difficulties by means of tests prepared for the purpose, and that the use of the tests without the accompanying remedial program is futile.

An examination of available practice and drill material in the field of arithmetic reveals two somewhat distinctive types and uses of such material. General practice exercises designed to simplify the first learning and to aid in maintaining a general mastery of skills are numerous and varying in type. They range from practice cards designed for repeated use to cheap practice tablets and comprehensive workbooks designed for drill and maintenance purposes. The arithmetic drill devices that have been constructed particularly for remedial purposes are not so numerous. Some idea of the organization of remedial material within a given field of arithmetic may be gained from the accompanying table, in which the units of remedial drill designed to correct difficulties in the manipulation of whole numbers are shown in their relation to the basic skills to be developed.

### Problem-solving exercises

Skill in the solution of verbal problems is difficult to develop because of its complexity. The remedial aspect of the field of problem-solving largely remains to be developed. Since a complete sampling of the complex mass of skills involved in problem-solving cannot be made, only a few of the most important skills are included in the accompanying proposals for teacher-made remedial material. Practice on the silent reading comprehension of verbal problems of varying degrees of complexity is suggested in this analysis. The selection of the facts given in the problem that are essential to its solution responds to practice. Skill in comprehending the real problem setting by determining what is called for in the problem is also



TABLE 35. Possible scope of drill units in whole numbers

For weaknesses in these Basic Skills	Prepare these types of Remedial Exercise Units
Basic Addition Facts	100 Addition Facts
Basic Subtraction Facts	100 Subtraction Facts; easy combinations, no carrying or borrowing
Basic Multiplication Facts	100 Multiplication Facts
Basic Division Facts	90 Division Facts
Higher Decade and Column Addition	Higher Decade Addition Facts
Carrying in Column Addition	Introducing easy carrying
Harder Subtraction	Harder subtraction combinations introducing borrowing or carrying
Addition Used in Harder Multiplication	Higher Decade Addition Facts
Carrying in addition used in Harder Multiplication	The 360 multiplication and addition combinations used in carrying in multiplication
Complete Process of Multiplication	Units introducing one-, two-, three-, and four-figure multipliers, and zero difficulties
Short Division with Carrying	The 360 short division combinations involving carrying
Addition, Subtraction, Multiplication used in Long Division	Previous units in these fields
Estimating Quotients	Units introducing estimation of apparent and non-apparent quotients
Complete Process of Long Division	Units introducing complete process of long division step by step

developed by practice on this type of exercises. Practice on the basic skill of choosing the correct process or combination of processes in the more complex problems is also suggested. Skill in the verification of the solution or the estimation of the most probable answer to the problem may be developed by special exercises. The complete set of problem exercises can be utilized as a means of unifying the skills involved in the complete process of problem-solving.

TABLE 36. Analysis of problem-solving

Steps in Problem-solving	Factors Underlying Problem-solving	Proposed Types of Drill
Comprehension	Vocabulary. Ability To Read Numerals. Ability To Read Rapidly. Ability To Comprehend. a. Follow directions. b. Make generalizations. c. Select potent elements. d. Discard irrelevancies. e. Determine problem setting as a unit. f. Determine the outcome of the problem. g. Grasp the significance of problem cues.	Multiple Choice. Comprehension. Exercises based on verbal problems.
Analysis and Organization	Selection of Potent Elements. Selection of Processes Involved. Determining What the Problem Calls for. Determining What Is Given in the Problem. Determining the Process Relationships.	What Is Given in Problems. Process Required in Solving Problems. What Is Called for in Problems. Problem Relationships.
Recognition	Choice of Procedure. Determining Problem Conditions. Determining the Purpose of the Problems. Determining Relevant Elements.	Processes Required in Solving Problems. What Is Called for in Problems. What Is Given in Problems.
Solution	Selection of Process. Organization of Processes in Order. Knowledge and Application of Combinations. Problem Relationships.	Processes Required in Solving Problems. Working on Problem Scales.
Verification	Probable Form of Answer. Probable Magnitude of Answer.	Estimation of Probable Answers.



The validity of any type of drill depends on the degree to which the sampling covers the fundamental skills and the degree to which the exercises themselves actually develop the skills they purport to develop. There are a number of places in which this complex chain may break. The task of diagnostic and remedial treatment is to locate and to repair quickly those links of the chain that have snapped under stress or have rusted out through lack of use.

Two further observations are important. First, while specific drill on some skill all by itself is often quite important, it must be accompanied by, if not preceded by, understanding of the total situation. Satisfactory performance on an isolated skill is not always matched by similar performance on the same skill when it operates in a more complex situation. Thus  $9 + 4$  may be an easy combination by itself, but it may not click at all when presented as  $7 + 2 + 4$ , where the 9 is unseen. From this it follows that after attention is paid to a specific breakdown the skill should also be practiced in the most complex situation in which it appears.

A second caution deals with a matter of policy. The need for remedial work of any kind and in any subject implies a failure at some point in the initial learning. Remedial work should be reduced as much as possible by making the first learning effective, by adequate review devices, and by the proper grade placement of pupils. A teacher should never be proud of the amount of remedial work he must do. However, he may be proud of his ability to direct it well when need for it arises. Obviously, preventive work based upon understanding is better teaching than remedial work.

### Workbooks in elementary-school mathematics

The importance of the workbook as a source of supplementary drill and maintenance material is indicated in a report of Johnson<sup>14</sup> showing that in a single year "the sale of workbooks, according to the best estimates, approximated thirty-seven million copies and the sales of workbook and test materials now consume nearly 25 per cent of the instructional budget." There is little reason to assume that these figures have declined since that report. Arithmetic, being a subject in which extensive practice and maintenance drill is necessary

<sup>14</sup> W. P. Johnson, "Then Came the Workbook." *Journal of Education*, 131:64-66; February 1948.

in order to fix and hold desirable habits and skills, doubtless comes in for its fair share of expenditures for workbooks.

There are widely varying opinions among teachers, administrators, and supervisors relative to the values of workbooks. Publishers naturally favor their use and in a published report recommend workbooks in reading, spelling, language, penmanship, and arithmetic. Some teachers hold that there is a purpose for every well-designed workbook. It certainly is true that there are workbooks for practically every purpose. Unfortunately, there is little experimental evidence on the instructional contributions of workbooks in the specific subject fields. Gray<sup>15</sup> pointed out the following general values of workbooks:

1. Timesaving. "The modern workbooks supply . . . devices that free the teacher of drudgery for the more creative aspects of teaching."
2. Meaningful practice. "Instead of endless copying, the student . . . has to think and to make decisions for each entry . . . to focus his attention on the essential points."
3. Adequate drill. "Each student gives a full recitation. . . Individual skill drill is important in many subjects."
4. Skill application. "... transforms the textbook into a functional experience."
5. Self-instruction. "... the child has his own ways of learning and should be left free to apply them."
6. Individual attention. "The modern workbook is self-diagnostic, revealing individual strengths and weaknesses. It provides drill where it is needed and enrichment activity for the better-equipped student."
7. Proportioned emphasis. "... the year's study is balanced so that the important topics will have attention."
8. Savings in cost. "Supplementary workbooks cost less than equivalent mimeographed materials."

It should be pointed out that undoubtedly there are those who would use these same eight points as the basis of their rejection of the workbook as a teaching aid. Basically the workbook is a textbook and as such may run the range of quality from poor to excellent. Moreover the efficiency of a given textbook or workbook depends to a large degree upon the manner in which it is used for instructional

<sup>15</sup> Albert Gray, "Lift the Workbook Cover." *Phi Delta Kappan*, 33:286-87; January 1952.



purposes by the individual teacher. The published evidence on the content of workbooks, the techniques of selecting workbooks in relation to the course of study, and the demonstrated instructional value of workbooks is very scattered. There is some evidence, not too elaborate nor convincing, that workbooks are helpful in mathematical instruction.

## 6 PREDICTION OF SUCCESS IN LATER MATHEMATICS

The growing tendency of high-school administrators and supervisors to utilize elementary-school records and test results as pre-registration information justifies some attention at this point to the problem of predicting success in secondary-school mathematics. The excessive difficulty which pupils encounter in first-year algebra has long impressed parents, teachers, and supervisors. In general, these difficulties are revealed in two ways: (1) in the high percentage of pupil failure in the subject, and (2) in the large amount of outside assistance required by pupils if failure is to be avoided.

Two somewhat different procedures have been followed as a basis for prognosis in mathematics beyond the elementary school. These are (1) the *learning* technique, in which the aptitude of the pupil for the subject is measured in terms of the speed and accuracy with which he is able to acquire skills and information in the new field and respond to objective tests over the newly learned material, and (2) the *inventory* technique, in which he reveals his aptitude in terms of reactions to specific exercises sampling into underlying skills upon which success in the subject depends.

The *Orleans Algebra Prognosis Test* is an aptitude test of the learning type. The test contains eleven simple lessons with a test on each covering fundamental principles and essential skills in learning algebra. An arithmetic and a summary test are also included. The *Iowa Algebra Aptitude Test*, Revised, is an example of the inventory type. The following four basic skill areas, none of which involves algebraic ability as such, have been shown to be highly related to achievement in first-year algebra: (1) arithmetic computations, (2) computations involving abstract concepts, (3) manipulation of numerical series, and (4) solution of problems involving dependence and variation. Both of these aptitude instruments may profitably be used at or near the end of the year preceding the introduction of first-year algebra. Pupils who score below specified critical points

on the test norms should probably be diverted into courses in general mathematics or into other fields of study where they are more likely to succeed.

### Topics for Discussion

1. Why does the content of elementary-school mathematics remain much more constant than does the content of the sciences?
2. What important new changes in the ultimate objectives of instruction in elementary-school mathematics have appeared within the last decade and how do these changes affect teaching procedures in this field?
3. What accounts for the fact that the field of arithmetic has been subjected to rather extensive analysis and intensive measurement?
4. Identify some of the more important of the specific skills in arithmetic that appear to lend themselves to measurement and remedial treatment.
5. To what extent does it appear justifiable to depend upon transfer to aid in the learning of arithmetic skills?
6. Illustrate some of the standardized test techniques that have been used in the measurement of general skills, problem-solving ability, and basic arithmetic concepts.
7. Show how the basic steps in problem-solving closely parallel the steps in the thinking process.
8. Describe the techniques proposed for the analysis of problem-solving abilities.
9. What are the chief differences in the principles underlying the three types of diagnostic tests in basic arithmetic skills?
10. Outline a survey, diagnostic, and remedial program in arithmetic, indicating where possible your first and second choices of material. Give reasons for your choices.
11. Summarize arguments for and against the use of arithmetic workbooks.
12. Outline a plan by which you could give helpful guidance to pupils of an eighth-grade class concerning the type of high-school mathematics they should elect.

### Selected References

*Arithmetic in General Education*. Sixteenth Yearbook of the National Council of Teachers of Mathematics. New York: Bureau of Publications, Teachers College, Columbia University, 1941.



- BROOM, M. E. *Educational Measurements in the Elementary School*. New York: McGraw-Hill Book Co., Inc., 1939. Chapter 8.
- BROWNELL, WILLIAM A. "Remedial Cases in Arithmetic." *Peabody Journal of Education*, 7:100-7; September 1929.
- BRUECKNER, LEO J. "Diagnosis in Arithmetic." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1935. Chapter 14.
- BRUECKNER, LEO J. "Significant Trends in Research in Diagnosis in Arithmetic." *Journal of Educational Research*, 33:460-62; February 1940.
- BRUECKNER, LEO J., AND GROSSNICKLE, FOSTER E. *How To Make Arithmetic Meaningful*. Philadelphia: John C. Winston Co., 1947. Chapter 10.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 504-18.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p. 281-302.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1939. p. 38-43.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. p. 419-38.
- BUSWELL, GUY T. "Contributions of Research to Special Methods: Elementary School Mathematics." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1938. p. 123-28.
- GLENNON, VINCENT J. "Testing Meanings in Arithmetic." *Arithmetic 1949*. Supplementary Educational Monographs, No. 70. Chicago: University of Chicago Press, 1949. p. 64-74.
- GREENE, CHARLES E., AND BUSWELL, GUY T. "Testing, Diagnosis, and Remedial Work in Arithmetic." *Report of the Society's Committee on Arithmetic*. Twenty-Ninth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1930. Chapter 5.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. p. 225-32.
- KNIGHT, FREDERICK B., chairman. *Report of the Society's Committee on Arithmetic*. Twenty-Ninth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1930.

- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. p. 120-33.
- PARRY, MARGARET E. "Arithmetic Tests." *Grade Teacher*, 65:70; March 1948.
- PARRY, MARGARET E. "Denominate Number Test." *Grade Teacher*, 67:84; March 1950.
- SPACHE, GEORGE. "A Test of Abilities in Arithmetic Reasoning." *Elementary School Journal*, 47:442-45; April 1947.
- SPITZER, HERBERT F. "Procedures and Techniques for Evaluating the Outcomes of Instruction in Arithmetic." *Elementary School Journal*, 49:21-31; September 1948.
- SPITZER, HERBERT F. *The Teaching of Arithmetic*. Boston: Houghton Mifflin Co., 1948. Chapter 12.
- SPITZER, HERBERT F. "Testing Instruments and Practices in Relation to Present Concepts of Teaching Arithmetic." *The Teaching of Arithmetic*. Fiftieth Yearbook of the National Society for the Study of Education, Part II. Chicago: University of Chicago Press, 1951. Chapter 10.
- SUEL TZ, BEN A. "Evaluation of Arithmetic Learnings." *National Elementary Principal*, 30:24-33; October 1950.
- SUEL TZ, BEN A. "The Measurement of Understandings and Judgments in Elementary-School Mathematics." *Mathematics Teacher*, 40:279-84; October 1947.
- SUEL TZ, BEN A. "Measuring the Newer Aspects of Functional Arithmetic." *Elementary School Journal*, 47:323-30; February 1947.
- SUEL TZ, BEN A., BOYNTON, HOLMES, AND SAUBLE, IRENE. "The Measurement of Understanding in Elementary-School Mathematics." *The Measurement of Understanding*. Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1946. Chapter 7.
- The Teaching of Arithmetic*. Tenth Yearbook of the National Council of Teachers of Mathematics. New York: Bureau of Publications, Teachers College, Columbia University, 1935.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapter 10.
- WILSON, GUY M. "Arithmetic." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 44-58.



## *Measuring and Evaluating in the Elementary Sciences*

THIS CHAPTER presents a discussion of the following points involved in the measurement and improvement of instruction in the elementary sciences :

- A. Objectives of the elementary sciences.
- B. Outcomes of the elementary sciences.
- C. Measurement in the sciences.
- D. Standardized tests in the elementary sciences.
- E. Testing methods in the elementary sciences.
- F. Informal objective testing of elementary science outcomes.

This chapter supplements the preceding one by furnishing a further discussion of the problems of measurement in the content subjects. It deals with the elementary science fields of nature study, hygiene, and general science.

Adequate science instruction in the elementary school should be expected as a matter of course in an age of science like the present. It is rather surprising to note, however, that progress in the selection and organization of science content and improvement in teaching and testing methods and materials in recent years have been relatively slight, in spite of the great practical value of science and its natural appeal to the curiosity and interests of elementary-school children. Science as an elementary-school subject has made but few contributions of experience or enrichment to the progress of education. Apparently most of the adult population have learned to make their

adjustments in this scientific age through experiences they have had outside of the elementary school.

## 1 SCOPE OF ELEMENTARY SCIENCES

### Objectives of elementary sciences

An examination of numerous modern sources fails to disclose any clear-cut and realistic statements of objectives for the elementary sciences. Aims and objectives appear to be more satisfactory for science in general and for the secondary school, however. The following list for science in general includes eight types of objectives but in each instance provides only a few of the illustrations given to show how the objectives can be attained.<sup>1</sup>

- A. *Functional information* or *facts* about such matters as:
  - 1. Our universe—earth, sun, moon, stars, weather, and climate.
  - 2. Living things—plants and animals.
  - 3. The human body—structure, functions, and care.
  - 4. Energy—sources, types of energy, machines.
- B. *Functional concepts*, such as:
  - 1. Space is vast.
  - 2. The earth is very old.
  - 3. All life has evolved from simpler forms.
- C. *Functional understanding* of principles, such as:
  - 1. All living things reproduce their kind.
  - 2. Energy can be changed from one form to another.
  - 3. All matter is composed of single elements or combinations of elements.
- D. *Instrumental skills*, such as:
  - 1. Read science content with understanding and satisfaction.
  - 2. Perform fundamental operations with reasonable accuracy.
  - 3. Read maps, graphs, charts, and tables and interpret them.
  - 4. Make accurate measurements, readings, titrations, etc.
- E. *Problem-solving skills*, such as ability to:
  - 1. Sense a problem.
  - 2. Define the problem.

<sup>1</sup> Victor H. Noll, chairman, "The Objectives of Science Instruction," *Science Education in American Schools*, Forty-Sixth Yearbook of the National Society for the Study of Education, Part I. University of Chicago Press, Chicago, 1947. p. 28-29. Quoted by permission of the Society.



3. Select the most likely hypothesis.
4. Test the hypothesis by experimental or other means.
5. Draw conclusions.

F. *Attitudes*, such as:

1. Open-mindedness—willingness to consider new facts.
2. Intellectual honesty—scientific integrity, unwillingness to compromise with truth as known.

G. *Appreciations*, such as:

1. Appreciation of the contributions of scientists.
2. Appreciation of basic cause-and-effect relationships.

H. *Interests*, such as:

1. Interest in some phase of science as a recreational activity or hobby.
2. Interest in science as a field for a vocation.

Noll <sup>2</sup> analyzed and classified science objectives from 130 sources and tabulated the frequency with which each type of objective was listed for the elementary school and the junior high school in the various sources. The objectives are listed below by types and in order of frequency of listing.

A. Knowledges

1. Knowledge of the principles and applications of science.
2. Knowledge leading to an understanding of the nature and organization of the environment.
3. Exploration to acquaint the pupil with science and to help him to orient himself with respect to the different sciences.
4. Preparation for further work in science and for college entrance.

B. Habits

1. Desirable habits of work and study.
2. Ability to do useful tasks.

C. Appreciations

1. Appreciation of the beauties of nature and of the commonplace.
2. Appreciation of the work of scientists.

D. Interests

1. Interest in science.
2. Interest in environment.

<sup>2</sup> Victor H. Noll, *The Teaching of Science in Elementary and Secondary Schools*. Longmans, Green and Co., New York, 1939. p. 7-10.

### E. Abilities

1. Ability to use the scientific method.
2. Ability to do useful tasks.

### F. Attitudes

1. Scientific attitude.

The major differences between the elementary school and the junior high school in emphasis on the various objectives was that knowledges were emphasized more in the junior high school whereas habits received greater attention at the elementary-school level. The emphasis upon the knowledge and habit objectives was approximately as great as that upon all four of the less tangible objectives combined.

## General outcomes of elementary sciences

The elementary sciences must be viewed from two rather specific points of view—for their immediate educational values for children of the elementary-school level, and for the background of preparation they afford for the later more intensive and specialized study of the sciences. Educational values of real significance will be attained if pupils, as a result of such instruction, acquire (1) the ability to use the scientific findings that apply in their experiences, (2) the ability to interpret natural phenomena in their environments, and (3) an appreciation of scientific attitude through understanding of and ability to use some of the methods of study that have been employed by scientists.<sup>3</sup>

The question of organization of courses arises here as it does in the social studies—whether the sciences should follow the traditional subject divisions or be integrated to produce a unified course of study. The tendency in the most progressive schools is toward unification. On the whole this movement has met with more general approval in the elementary grades than it has at the secondary-school or college levels.

Regardless of the desirability of developing an integrated course of study, most of the elementary-school science now taught is presented

<sup>3</sup> S. Ralph Powers, "The Plan of the Public Schools and the Program of Science Teaching." *A Program for Science Teaching*, Thirty-First Yearbook of the National Society for the Study of Education, Part I. Public School Publishing Co., Bloomington, Ill., 1932. p. 10.



in the form of separate courses in nature study, physiology, and general science. The scope of each of these is discussed here without reference to their possible integration.

*Nature study.* The direct needs of life to which nature study contributes are of three kinds—economic, hygienic, and appreciative. Knowledge concerning how plants and animals serve human needs involving soils, climatic conditions and effects, tillage, control of pests, plant and animal foods, and means of preserving plant and animal products is important. Much of this knowledge is biological. There is also much need for knowledge of the physical sciences, such as the simpler operations and principles of physics and chemistry, in connection with food, clothing, shelter, transportation, and other everyday problems.

The appreciative needs of nature study are of two kinds—aesthetic and intellectual. The aesthetic needs grow out of interests in the beauties of nature. The revelations of beauty in plant and animal forms, in land and water formations, and in earth and sky by day and night furnish much enjoyment. Furthermore, the cultivation of these interests affords purpose to many recreational activities. Intellectual needs resulting from the natural curiosity of children in how and why the forces of nature operate as they do are satisfied by the study of nature. This curiosity may be developed by cultivation into a permanent interest in nature and science.

*Physiology and hygiene.* The development of proper habits in caring for the body requires some knowledge of the structure and use of its parts. The general structure of the teeth, the skin, the nails and hair, the eyes, the ears, the nose, the throat, and the mouth should be known for the contribution such knowledge makes toward keeping them all in a healthful condition. A general knowledge of the digestive organs, lungs, circulatory system, organs of excretion and sex, and the nervous system is useful in keeping these organs healthfully at work. This body of knowledge and the health habits developed with it constitute the hygienic aspects of nature study. As health knowledge tests will be considered in Chapter 21 of this volume, physiology and hygiene tests are treated here only to the extent to which they enter into general tests for the elementary sciences.

*General science.* Most intelligent adjustments, as distinguished from those that are purely accidental, impulsive, or habitual, are dependent upon scientific procedures. Everyone is called upon to

make such responses in connection with his home, his neighborhood, his vocation, his civic duties, and his leisure. He is frequently confronted with a need for some special knowledge of health control, mechanics, chemistry, physics, biology, or plant and animal life. At most every hour of the day the individual is in the midst of the influence of mechanical and scientific appliances. For their operation, maintenance, adjustment, and repair, and as a protection from their dangers, he needs information and first-hand experience of the type obtained in general science.

## 2 MEASUREMENT IN ELEMENTARY SCIENCES

### Difficulties in constructing science tests

The construction of science tests should apparently be relatively simple, since the content of science is quite tangible. However, difficulties of a degree no less marked than in the other content subjects are encountered. There is the same lack of agreement on the content of the course of study and its organization that is found in the social sciences. Controversies about the importance of facts as contrasted with emphasis on relationships and problem-solving are still somewhat in evidence in science teaching, although science teachers have increasingly of late given attention to the more intangible outcomes of instruction. There is very little objective evidence on what particular skills and principles, or what elements in and safeguards to scientific thinking, are of most importance or can best be imparted in the elementary-school sciences. The typical science course apparently attempts to accomplish little more than to give a knowledge of the names of a few of the common animals, plants, and physical objects, and an acquaintance with a few of the simpler natural phenomena, without any very definite purpose appearing to justify the accumulation of such information.

Real evidences of accomplishment in the sciences are to be found in the development and the direction of pupils' interests, attitudes, appreciations, skills, habits, and actions in these fields. The ideal way to determine the changes that are effected in the pupil as a result of studying a unit in science would be to measure the increment of desirable activities that he can and does perform as a result of this study. However, only a few attempts to devise tests for such a purpose have so far been made.



## Measurable outcomes of science

Five major types of measurable qualities are designated in the sciences.

*Knowledges.* Most tests in science tend to overemphasize information and knowledge as the goal of study. It is too often assumed that knowledge is a positive index of satisfactory modes of adjustment. This assumption, of course, is only partially defensible. Merely to know is no assurance of subsequent proper reaction. But insofar as knowledge is essential to adjustment its proper worth should not be discounted. Accordingly, measurement of the pupil's knowledge of scientific facts is to that extent valid and defensible.

*Skills.* Although laboratory and other science skills are not as much involved in the elementary-school as in the secondary-school sciences, the degree to which pupils attain the desired skills can be measured readily. Performance rather than paper-and-pencil tests are often demanded in such situations. Since performance tests are treated in Chapter 8, the measurement of skill outcomes in the sciences receives little attention in this chapter.

*Concepts and understandings.* Facts in science are the vehicles for thought. The understanding of the relationships of facts and of generalized ideas is deemed most important. It is these generalized ideas that pupils should attain in their study of science. Tests should, therefore, so far as possible, measure the relational aspects of science, and do succeed in this aim to a reasonable degree.

*Applications.* Problem-solving tests in science call for the application of knowledge and may demand one or more types of scientific thinking. Similarly, test items that involve the interpretation of new situations demand more than mere recall and, thus, are measures of ability to use scientific knowledge or judgment. Such test items should find a more extensive place in testing procedures than they have thus far been given.

*Attitudes and interests.* Some progress has been made in the measurement of pupils' attitudes toward and interests in science material and phenomena. The task is a difficult one, because as yet the interests and the attitudes deemed desirable have not been defined very clearly. Furthermore, it is not too clear how pupils should be tested, or what should be the content of the test that will reveal their possession of the desired interests and attitudes in a

dynamic sense. Some significant attempts have been made in the measurement of scientific attitudes, however.

### 3 STANDARDIZED TESTS IN ELEMENTARY SCIENCES

#### Standardized tests in course areas

The number and variety of standardized elementary science tests is not great. Tests for the intermediate grades are found mainly as parts of achievement test batteries, and these parts are seldom available in separate booklets. Only one nature study test is known to the writers, although the *Modern School Achievement Test* includes some nature study materials in its elementary science section. No physiology and hygiene tests are known, but attention is devoted to that subject in the elementary science sections of the *Stanford Achievement Test* and the *Progressive Tests of Social and Related Sciences*.

Standardized tests in nature study and probably in physiology and hygiene are dependent upon (1) a more universal agreement on their aims and purposes, (2) more representative criteria for course of study content, and (3) a more definite identification of their minimum essentials. A trend seems apparent, however, toward the merging of both of these courses with the somewhat broader course in elementary science.

Significant development of a unified course in science continues in Grades 7, 8, and 9. The general science course has not been widely accepted in the senior-high-school grades, with the result that the separate courses of biology, chemistry, and physics are still taught in most high schools. Tests are available for the general science course typically given in the ninth grade. Test batteries recently published have general or natural science sections for the junior-high-school grades and at least two of them provide separate-booklet editions for the natural sciences.

#### Standardized test methods

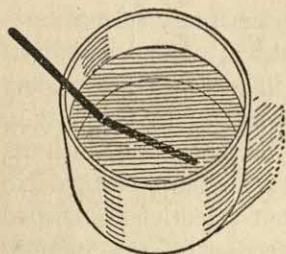
Sample items illustrative of the manner in which various objective item forms are used in elementary science testing are presented here. The student should utilize the sample items together with the bibliog-



raphy at the end of the chapter for information concerning standardized tests as well as for suggestions on types of informal objective items suitable for use in the elementary sciences.

*Simple recall items.* The following sample shows the manner in which simple recall items can be used with pictorial representation.

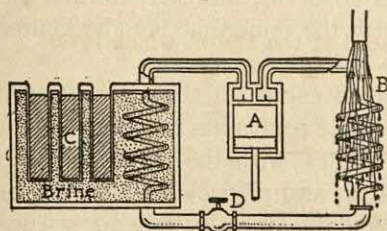
Sample A.<sup>4</sup>



- a* This diagram illustrates the facts of the ..... of light. *a*
- b* The amount of apparent bending of the stick depends upon the ..... of the liquid. *b*

*Completion items.* Simple recall and completion items differ only slightly in form and not at all in the nature of the pupil's response. The first two items below are of the simple recall or sentence completion type and the third is of completion form.

Sample B.<sup>5</sup>



In this drawing of an artificial ice plant:

- a* The freezing vats are located at ..... *a*
- b* The condensing pump is located at ..... *b*
- c* The principle involved in the manufacture of artificial ice by ..... is that the liquid turns into a ..... when the pressure is removed and, in so doing, it takes up ..... from the brine, which in turn ..... the temperature of the ..... in the freezing vats. *c*

cial ice by ..... is that the liquid turns into a ..... when the pressure is removed and, in so doing, it takes up ..... from the brine, which in turn ..... the temperature of the ..... in the freezing vats.

*True-false items.* The following sample of true-false items illustrates one of the few applications of this item type in elementary science tests.

<sup>4</sup> Giles M. Ruch and Herbert E. Popenoe, *Ruch-Popenoe General Science Test*. Published by World Book Co., 1923.

<sup>5</sup> *Ibid.*

Sample C.<sup>6</sup>

1. The common garter snake is poisonous. T F 1
2. The reason that one cannot see stars in the daytime is that the sun shines so brightly. T F 2
3. Air near the ceiling of a room is always warmer than air near the floor. T F 3

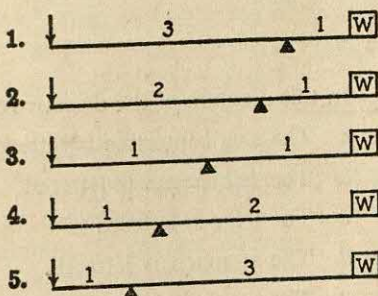
*Multiple-choice items.* By far the most popular item form in elementary science tests, the multiple-choice type, is used in several different adaptations. Samples D to F show sample items of the common type, an item based on diagrams, and an item based on a passage to be read.

Sample D.<sup>7</sup>

- <sup>1</sup> A "closed season" protects — 1 wild life 2 swimmers 3 hunters 4 travelers 1 2 3 4  
1 2 3 4
- <sup>2</sup> Many insect larvae are eaten by — 5 bees 6 flies 7 birds 8 worms..... 1 2 3 4  
1 2 3 4
- <sup>3</sup> A plant that grows from a bulb is the — 1 carrot 2 tomato 3 lettuce 4 onion 1 2 3 4  
1 2 3 4

Sample E.<sup>8</sup>

49. If a 500-pound weight is placed at the arrow, which lever will lift the 60-pound weight W the highest?



<sup>6</sup> Georgia S. Adams and John A. Sexson, *Progressive Tests in Social and Related Sciences, Test 6, Elementary Science*, Elementary. Published by California Test Bureau, 1946.

<sup>7</sup> Truman L. Kelley and others, *Stanford Achievement Test, Science*, Advanced battery, Form J. Published by World Book Co., 1953.

<sup>8</sup> John G. Read, *Read General Science Test*. Published by World Book Co., 1950.



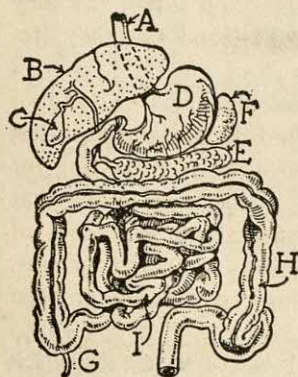
Sample F.<sup>9</sup>

Water takes up oxygen from the air in varying amounts. Cold water will take up small quantities of oxygen while warm water takes up almost none. Running water will dissolve (that is, take up) more oxygen than standing water. Water in which plants are growing contains much oxygen because the green plants give off oxygen in the process of photosynthesis. When there is not enough light for plants to manufacture food, they do not give off oxygen but consume it in respiration. Water animals also use oxygen in respiration so that the amount of oxygen found in water is always changing. The oxygen content of an aquarium changes from day to day and from hour to hour and is different even at different levels in the aquarium.

## 13. Standing water takes up

- 13-1 more oxygen than running water.  
 13-2 as much oxygen as running water.  
 13-3 less oxygen than running water.  
 13-4 a great deal of oxygen.  
 13-5 no oxygen. . . . . 13( )

*Matching exercises:* Two samples of the matching test are given below. The first, illustrating an identification test, requires the matching of parts of the digestive tract and their pictorial representation, and the second is a matching unit having some elements in common with the multiple-choice form.

Sample G.<sup>10</sup>

In this diagram of the digestive tract:

- |   |   |   |
|---|---|---|
| a | The small intestine is lettered . . . . | a |
| b | The esophagus is lettered . . . .       | b |
| c | The liver is lettered . . . .           | c |
| d | The stomach is lettered . . . .         | d |
| e | The pancreas is lettered . . . . .      | e |

<sup>9</sup> John G. Zimmerman and Richard E. Watson, *Cooperative Science Test for Grades 7, 8, and 9*, Form R. Published by Cooperative Test Service, 1941.

<sup>10</sup> Ruch and Popenoe, *op. cit.*

Sample H.<sup>11</sup>

1 Heart	1. Circulatory system . . . . .	1( )
2 Kidney	2. Excretory system . . . . .	2( )
3 Lung	3. Nervous system . . . . .	3( )
4 Stomach	7. Laws of biological inheritance . . . . .	7( )
5 Spinal cord	8. Germ theory of disease . . . . .	8( )
1 Mendel	9. Plant breeding . . . . .	9( )
2 Pasteur		
3 Hooke		
4 Burbank		
5 Carrel		

## 4 TESTING IN ELEMENTARY SCIENCES

Objective items of the types illustrated above have been used quite widely by informal objective test makers in the evaluation of the more intangible outcomes of science instruction.<sup>12</sup> Furthermore, rather complex adaptations of the common item forms have also been used. There seems excellent reason to believe that much of the most significant recent testing in the science field has been done by informal objective testing methods. Perhaps three major reasons why these techniques have not yet entered widely into the standardized testing field below the high-school and college levels are that (1) they have been applied in the main to situations requiring a variety of applications of science knowledges and skills, (2) they are rather difficult to construct and standardize, and (3) they frequently run to considerable length.

Space does not permit many illustrations here of this type of approach to the measurement of scientific knowledges and abilities. Although some of the illustrations given above are from testing for the junior or the senior high-school levels, it does not follow that these techniques are applicable only to instructional outcomes of greater complexity than those of the elementary school. It means, rather, that the most significant work of this type has so far been

<sup>11</sup> O. E. Underhill and S. R. Powers, *Cooperative General Science Test*, Form Q. Published by Cooperative Test Service, 1940.

<sup>12</sup> For example, see Louis M. Heil and others, "The Measurement of Understanding in Science." *The Measurement of Understanding*, Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. University of Chicago Press, Chicago, 1946. Chapter 6.



done at the high-school and college levels. The same techniques can be, and in some cases have been, adapted to the elementary sciences. Several of the evaluative and interpretive tests illustrated in Chapter 9 apply to elementary science.

### Measurement of broad instructional outcomes

An illustration of the application of evaluative techniques to the measurement of broad instructional outcomes of elementary science was given by LuPone<sup>13</sup> for a unit on machines and their applications. The following chart shows the relationships LuPone established between pupil outcomes, tools used in evaluating the outcomes, and illustrations of how the tools might be applied. Several illustrations of item types that cannot be reproduced here are also given in the reference. Observation of the tools used in the paralleling areas of evaluation and of the typical illustrations given in the chart should indicate to the student the possibilities of measurement employing a variety of techniques.

EVALUATION CHART

<i>Areas of Evaluation</i>	<i>Tools Used</i>	<i>Typical Illustrations</i>
CONCEPTS		
1. Our ways of living are affected by the use of machines.	Classroom tests.	Problems.
2. Man's intelligent endeavor has been a factor in our present civilization.	Work sheets.	
3. Our society is affected by inventions.	Parent interviews.	
4. Machine power is more efficient than man power.	Anecdotal records of pupil behavior based upon teacher observation.	
5. The era of machines has implications which are social.	Pupil logs or diary of the child's experience during the area of work.	

<sup>13</sup>O. J. LuPone, "Evaluating the Intangibles in Elementary Science." *School Science and Mathematics*, 39:754-59; November 1939.

## EVALUATION CHART (cont.)

<i>Areas of Evaluation</i>	<i>Tools Used</i>	<i>Typical Illustrations</i>
6. Human energy can be conserved by use of machines.		
KNOWLEDGES		
1. A knowledge that machines can do work more quickly, more easily, better, than man power.	Performance tests. Anecdotal records.	Experiments with simple machines.
2. Machines are a combination of two or more simple machines.	Pupil logs and pupil diaries.	Problems.
3. Machines give us more leisure time for recreation.		
ATTITUDES AND APPRECIATIONS		
1. Pupils have respect for people who develop more efficient ways of doing.	Comparison of pretests and final tests covering attitudes and appreciation.	Statements about the unit of work by which the child can express what he believes.
2. An appreciation that the standard of living is higher because of machines.	Classroom tests.	
3. A recognition that society changes through the effect of inventions.	Observations.	
OVERT BEHAVIOR		
1. A desire to visit machines at work.	Excursions.	A visit to a nearby project under construction.



EVALUATION CHART (*cont.*)

<i>Areas of Evaluation</i>	<i>Tools Used</i>	<i>Typical Illustrations</i>
2. A desire to read and send for material about machines.	Anecdotal records.	
3. A desire to use simple machines.	Writing for source materials.	
<b>SKILLS</b>		
1. Use and construct simple machines.	Performance tests.	The construction of simple machines.
2. Organization of materials.	A classroom test based upon skills devised by the teacher.	
3. Manipulation of apparatus.		

An informal, semi-objective test for teaching more than for testing purposes was devised by Davis<sup>14</sup> for use in eighth- or ninth-grade science courses in the measurement of other than largely factual instructional outcomes. The following reproduction of the instructions to pupils and of the first paragraph of the selection to be read and evaluated by the pupils will serve to show the nature of the instrument.

## TO THE PUPIL

Here is a test which I think you will find quite different from any you have ever taken. It is a story about Johnny Jones. He was quite an active boy, but sometimes he was a poor scientist. Some of his friends and the members of his family may not have been good scientists either. Whenever you find something in the story which does not agree with what you think good science means, put a pair of parentheses ( ) around the sentence or part of a sentence where you find this. Next, at the border of the paper beside the error, write in the correct letter from the following list:

S means that Johnny or some one else was superstitious.

D means that what was being done or had been done was dangerous.

<sup>14</sup> Warren M. Davis, "A Science Test Designed To Teach and Measure Outcomes Other Than Memorization of Factual Information." *Science Education*, 23:371-72; December 1939.

O means that statements are being taken or have been taken for truth without any proof being offered.

J means something unscientific for reasons other than S, D or O. If you use the letter J be prepared to tell the class what was wrong with the story at the point where you use this letter.

Now go on with the story.

### JOHNNY'S DAY

Johnny Jones woke from a sound sleep one morning and noticed that the sun was already shining in his window. Without looking where he was going he jumped to the floor and started gathering up his articles of clothing to put them on. Suddenly he stopped and said, "Shucks, it's Saturday, no need for me to hurry. But it might just as well be a school day," he went on as he looked out of the window, "it's sure to rain today. Old man Smith said this was a wet moon."

The remaining parts of the selection, running to perhaps 1100 words, included many additional evidences of behavior or reasoning illustrative of the types of situations covered by the S, D, O, and J methods of marking the selection. One point of credit was assigned for each pair of parentheses placed approximately in the correct position and an additional point of credit was assigned for each pair of parentheses accompanied by the proper identifying letter in this semi-objective test.

### Measurement of scientific attitude

Noll listed the following six abilities as essential to the scientific attitude: (1) accuracy in all operations—calculation, observation, and report, (2) intellectual honesty, (3) open-mindedness, (4) the habit of looking for natural causes, (5) the habit of suspended judgment, and (6) the habit of criticism.<sup>15</sup> Although he admitted that other habits might be included in such a list, he stated that a person who met all of the conditions listed above would possess the scientific attitude and would also be highly unique even in this scientific era.

<sup>15</sup> Noll, *op. cit.* p. 25-26.



Suggestions concerning how each of these six essentials of scientific attitude can be measured informally were also presented by Noll.<sup>16</sup> Some of his illustrations are reproduced to show techniques useful in measuring scientific attitude.

(1) Accuracy in calculation—arithmetic examples.

Accuracy of observation and report—questioning a pupil concerning the characteristics of an animal picture, plant, or diagram.

(2) Intellectual honesty.

T F When a pupil makes a poor mark in an examination it is usually because he is not well or he was up late the night before.

T F It is perfectly justifiable not to pay one's fare on a bus or street car if the conductor doesn't come around to collect it.

(3) Open-mindedness.

T F All Indians are dirty.

T F College professors as a rule would be failures in any line of work but teaching.

(4) Cause and effect relationships.

T F Finding a horseshoe means that one will have good luck.

T F Giraffes have such long necks because through many generations they have been stretched a little longer each time.

(5) Suspended judgment.

T F My neighbor is away from home most of the time. He must be a traveling salesman.

T F Mr. Jones bought a new car last week. He must have had an increase in salary.

(6) Criticism.

T F One can always accept as true what is printed in a book.

T F If my science teacher says a thing is so, it must be so.

Another approach to the measurement of similar types of outcomes is that by Davis, who presented the directions to pupils and a few sample items from a test for measuring knowledge of cause and effect relationships.<sup>17</sup> Students were asked to indicate by the use of the appropriate letter of the following

<sup>16</sup> *Ibid.* p. 34-37.

<sup>17</sup> Ira C. Davis, "The Measurement of Scientific Attitudes." *Science Education*, 19:117-22; October 1935.

- A—If the first occurrence is practically the sole cause of the second.
- B—If the first occurrence is one of a number of the important contributing causes of the second.
- C—If the first occurrence contributes only slightly to the second.
- D—If both occurrences are results of the same general cause or causes.
- E—If the first occurrence bears no causal relationship to the second.

their reactions to such items as these

1. The sun shines on the earth; the earth is warm.
2. A boy often picked up toads; the boy had warts on his hands.
3. The light of lightning; the accompanying thunder.
4. The ignition switch of an auto is turned on; the motor starts running.
5. A rising column of air was cooled; a cloud formed.

Davis also gave similar illustrations from a test designed to measure ability to distinguish between fact and theory.<sup>18</sup> The appropriate letter from this list

- A—Some are statements of well established facts which are always true.
- B—Others may be statements of well established theories which are generally accepted.
- C—Others may be statements of theories which are questioned by some (many) authorities.
- D—Others may be statements of popular beliefs which are not supported by evidence.

was to be used in responding to each of these statements

1. A disease is a punishment for some particular moral wrong.
2. Air is composed of molecules.
3. The pressure in water varies with the depth.
4. Heating the molecules in air increases their speed.
5. A high forehead indicates high intelligence.

### Measurement of superstitious beliefs

Zapf presented a technique for measuring the manner in which pupils actually behave in situations to which well-known superstitious beliefs apply.<sup>19</sup> Pupils were placed in a closed room, where they opened boxes in which were found directions for their subsequent

<sup>18</sup> *Ibid.*

<sup>19</sup> Rosalind M. Zapf, "Superstitious Beliefs." *School Science and Mathematics*, 39:54-62; January 1939.



action asking that they go contrary to widely held superstitious beliefs. The extent to which they performed the actions was taken as an indication of the degree to which they were not governed in their behavior by these beliefs. Such situations as breaking a mirror, walking under a ladder, and opening an umbrella indoors were among the twelve used in the test. Although all thirty-two pupils tested in these situations had previously indicated that they did not believe in the superstitions, only two pupils went contrary to all twelve superstitions and two pupils acted superstitiously in five of the twelve situations.

### Controlled observation

A controlled observation procedure for use in elementary-school science was worked out by West.<sup>20</sup> He devised a tabulation sheet and observation procedures of too great complexity for presentation here for use in the classroom evaluation of the dynamic and the performance factors of pupil behavior. Inquiry, critical-mindedness, open-mindedness, generalizing, recognition of achievements of thinking, scientific problem-attack, recognition of interpretations of natural phenomena, and cause-and-effect relationships were listed as dynamic factors, while responsibility, voluntary activity, initiative, application of experience, self-appraisal, resourcefulness, skills, special abilities, work habits, and miscellaneous were listed as performance factors. His recommendation was that this objective type of observational procedure be used to supplement but not to supplant the measurement techniques in common use in the classroom.

## 5 DIAGNOSIS AND REMEDIAL TEACHING IN ELEMENTARY SCIENCES

### Limitations of diagnostic and remedial techniques in science

Diagnostic procedures and remedial work in the field of science instruction are not highly developed. While certain of the available tests may show pupils to be deficient in some specific phase of science information, the majority of such tests do not point out the

<sup>20</sup> Joe Y. West, *A Technique for Appraising Certain Observable Behavior in Science in Elementary Schools*. Contributions to Education, No. 728. Teachers College, Columbia University, New York, 1937.

causes of the deficiencies. Practically all that can be done by way of diagnosis is in connection with certain skills that appear to be basic to the study of science.

The study of science involves the comprehension of a language peculiar to the subject. Reading of scientific content is apt to be difficult. Thus, poor reading ability may form the basis of poor accomplishment in the subject. Diagnosis of reading abilities of the work-study type, accompanied by remedial instruction designed to overcome the weaknesses revealed, is one of the prerequisites to satisfactory progress in the study of the sciences. Laboratory work may call for many new abilities and techniques.

### Future of diagnosis in science

There is considerable promise for the future of diagnosis and remediation in the sciences through further development of the evaluation techniques illustrated in the preceding section of this chapter. The attempt so far has been more upon the construction of valid evaluation procedures for the less tangible outcomes of instruction than upon diagnostic values of the techniques. The writers believe, however, that constructive diagnostic and remedial procedures may well grow out of this new approach to the measurement of ability in the sciences.

### Topics for Discussion

1. Why is objective measurement in the sciences not highly developed?
2. Enumerate and evaluate the aims of science education.
3. In your opinion is the need for a unified course in science in the intermediate grades any less serious than it is in the social studies?
4. What are the most important measurable outcomes of instruction in science?
5. Examine the possibilities of measuring the major outcomes in science and specify a number of techniques for each. Would such a list parallel the types found in the social studies?
6. What appears to be the present tendency in nature study and physiology and hygiene in the elementary science field?
7. Suggest some of the objective item types useful in science testing and illustrate them with items of your construction.
8. Discuss and evaluate the informal objective test approaches to the measurement of some of the more intangible outcomes of science instruction.



## Selected References

- ARNOLD, DWIGHT L. "Testing Ability To Use Data in the Fifth and Sixth Grades." *Educational Research Bulletin*, 17:255-59, 278; December 7, 1938.
- BUCKINGHAM, GUY E., AND LEE, RICHARD E. "A Technique for Testing Unified Concepts in Science." *Journal of Educational Research*, 30:20-27; September 1936.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 621-46.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p. 380-404.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938. p. 49-53, 62-65, 121-23, 143-46.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. p. 572-604.
- CURTIS, FRANCIS D. "Diagnosis and Remedial Treatment in the Field of Science." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1935. Chapter 16.
- CURTIS, FRANCIS D. *A Second Digest of Investigations in the Teaching of Science*. Philadelphia: P. Blakiston's Son and Co., 1931.
- CURTIS, FRANCIS D. *A Third Digest of Investigations in the Teaching of Science*. Philadelphia: P. Blakiston's Son and Co., 1939.
- DOWNING, ELLIOT R. "Some Results of a Test on Scientific Thinking." *Science Education*, 20:121-28; October 1936.
- FRUTCHEY, FRED P. "Evaluation in Elementary School Science." *Educational Method*, 16:422-26; May 1937.
- FRUTCHEY, FRED P., AND TYLER, RALPH W. "Examinations in the Natural Sciences." *The Construction and Use of Achievement Examinations*. Boston: Houghton Mifflin Co., 1936. Chapter 5.
- HEIL, LOUIS M., AND OTHERS. "The Measurement of Understanding in Science." *The Measurement of Understanding*. Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1946. Chapter 6.
- JOHNSON, PHILIP G. "Some Developments in Science Teaching and Testing." *School Science and Mathematics*, 50:187-99; March 1950.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. Chapter 10.
- LUPONE, O. J. "Evaluating the Intangibles in Elementary Science." *School Science and Mathematics*, 39:754-59; November 1939.

- NOLL, VICTOR H. *The Teaching of Science in Elementary and Secondary Schools*. New York: Longmans, Green and Co., 1939. Chapter 8.
- NOLL, VICTOR H., chairman. "Judging the Results of Instruction in Elementary Science." *Science Education in American Schools*. Forty-Sixth Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1947. Chapter 8.
- POWERS, SAMUEL R. "Science Education." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 1133-45.
- READ, JOHN G. "A Non-Verbal Test of the Ability To Use the Scientific Method as a Pattern for Thinking." *Science Education*, 33:361-66; December 1949.
- REINER, WILLIAM B. "Evaluating Ability To Recognize Degrees of Cause and Effect Relationships." *Science Education*, 34:15-28; February 1950.
- VERDUIN, JACOB. "An Open-Book Objective Examination for Science Courses." *School Science and Mathematics*, 50:213-21; March 1950.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapter 16.
- WEBB, SAM C. "A Generalized Scale for Measuring Interest in Science Subjects." *Educational and Psychological Measurement*, 11:456-69; Autumn 1951.
- WEITZMAN, ELLIS, AND McNAMARA, WALTER J. *Constructing Classroom Examinations*. Chicago: Science Research Associates, 1949. p. 46-47.
- WEST, JOE Y. *A Technique for Appraising Certain Observable Behavior in Science in Elementary Schools*. Contributions to Education, No. 728. New York: Bureau of Publications, Teachers College, Columbia University, 1937.



## *Measuring and Evaluating in the Fine Arts*

THE FOLLOWING possibilities of measurement of aptitudes and achievement in the fine arts are discussed in this chapter:

- A. Social and educational significance of the arts.
- B. Educational emphasis on music and art.
- C. Basic elements of musical talent.
- D. Measurement of musical accomplishment.
- E. Measurement of art appreciation.
- F. Measurement of artistic ability.

The objective measurement of aptitudes and achievement in the fine arts is a relatively recent accomplishment—so recent, in fact, that there is still an echo of protest from a small group of artists that artistic production does not lend itself to objective evaluation. In spite of this feeling, however, much progress has been made in these fields of measurement. This is as it should be, for certainly in these cultural subjects is to be found much of the best that the educational program affords. With the trend of recent years in the direction of greater individual leisure for the cultural pursuits, the need for a better understanding of the content, aims, and methodology of these artistic subjects is greater than ever before.

There is perhaps a certain advantage in the fact that developments in the measurement of the fine arts have taken place somewhat slowly. In general, research techniques have improved, with the net

result that the problems of measurement in these fields have been more critically analyzed and attacked with more refined instruments. The careful research of such critical workers as Carl Seashore, Schoen, Stanton, Kwalwasser, and Dykema—to name only a few in the psychology and pedagogy of music—and the work of Thorndike, Ayer, Meier, Manuel, Winslow, and Whitford—as an incomplete sampling of important names in the field of art—are evidences of the influence of this scientific point of view.

## 1 MEASURABLE QUALITIES IN MUSIC

### Musical talent and achievement

Measurement in music takes two major lines of approach. The first is the determination of basic aptitudes. Here, as in other subjects, the techniques and instruments used are psychological. Such instruments have been mentioned previously in this volume as tests of specialized intelligence, since they have to do with the determination of tendencies to respond in certain ways to specific types of musical stimuli. Accomplishment in music depends to such a large degree upon the existence of aptitude that this phase of measurement must be given primary emphasis. The mere existence of aptitude in music is in no sense an index to musical accomplishment, however. The second approach to the problem is pedagogical and is based upon the use of achievement tests for the threefold purpose of measuring the knowledges, skills, and appreciative aspects acquired as a result of training. As Kwalwasser<sup>1</sup> pointed out: "Regardless of the talent possessed, one must have the will to succeed or little is attained . . . There are a vast number of reasons why an individual of superior endowment may realize but a very small return on his native musicianship."

### Major aims and outcomes of music education

The statement of aims and outcomes of music education that appeared in the 1921 report of the *Educational Council of the Music Supervisors National Conference*<sup>2</sup> has not been greatly improved

<sup>1</sup> Jacob Kwalwasser, *Tests and Measurements in Music*. C. C. Birchard Co., Boston, 1927.

<sup>2</sup> Report of Educational Council of the Music Supervisors' National Conference. National Education Association, Washington, D. C., 1921.



upon since that time. The statements of instructional outcomes and the recommended standard course of study in music performed two very admirable functions: (1) they provided tangible goals for teachers and supervisors, and (2) they provided defensible criteria for the validation of tests of musical accomplishment. The student in this field will do well to investigate this report.

A more recent and very useful statement of the major goals of elementary-school music is that presented by Brooks and Brown.<sup>3</sup> It is believed that this general summary of elementary-school music instructional goals also affords a very useful basis for the validation of improved tests in the fields of music information, accomplishment, and aptitudes. Fifteen of these major goals are reproduced here and are summarized with minor modifications under the following seven practical categories:

1. In Song Singing

Ability to . . . use the voice to express and convey musical meaning in free, spontaneous, and beautiful song singing and with artistic interpretation.

2. In Chorus

Ability and disposition to associate with others . . . in joint rendering of music in chorus singing. . . .

3. In Appreciation and Its Background

a. Discrimination and taste in music with evidence of preference for that which has excellence and worth.

b. Sensitiveness to ordered perfection of structure and design in music both in song and in instrumental compositions and realization of aesthetic satisfaction in the beauty, appropriateness, and adequacy thus seen and expressed.

c. Integrated volitional structure in personality with reference to selection in music, which leads to the choice and use of music which has high excellence in contrast to that which is inferior in quality. . . .

d. Understanding of some phases of the development of music and some insight into the essential nature and meaning of music and the forces and influences that have produced it—including knowledge about musicians, familiarity with compositions, acquaintance with instruments and how they developed—to an extent and on a level appropriate to children of elementary-school age. . . .

<sup>3</sup> Marian Brooks and Harry A. Brown, *Music Education in the Elementary School*. American Book Co., New York, 1946. p. 114-16.

## 4. In Instrumental Music

- a. Ability to use instruments as a means of musical expression and with satisfaction in such experience.
- b. Ability to handle with manipulative skill such musical instruments as are used.

## 5. In Creative Music

Ability to use individual originality and personal initiative in interpreting, using, and creating music.

## 6. In Connection with the Musical Score

- a. Ability to read musical meaning fluently from the printed score.
- b. Ability to use musical notation to express or record musical meaning.
- c. Understanding of selected phases of the theory of music to an extent and on a level appropriate to elementary-school pupils, as essentially a functional approach to music literature and as a means toward a broader interpretation, including such elements of musical structure as accent, measure, phrase and period, scale and chord building, lines and spaces, key signatures.

## 7. In Connection with More Than One Phase of Music Education

- a. Complete freedom from inhibitions arising from focal attending to mechanical processes, accomplished by the development of an habitual-response pattern that releases conscious attention from the mechanics and structural details and permits complete absorption in getting or expressing meaning particularly (1) in singing, absence of conscious attention to the manner and the acts involved in utterance, (2) in interpreting the musical score in reading music, freedom from focal consciousness of the structural elements involved in the symbol perception necessary in gaining musical meaning, (3) in instrumental music, absence of focal attention to the finger manipulations and other physical movements connected with handling and managing the instrument.
- b. Ability to sense and feel the movement resident in music and to express it in bodily motion in some appropriate manner. . . .
- c. Growth toward possession of music as a social institution on a child's level of comprehension and participation; manifested by (1) children thinking and feeling together in groups and co-operating in collective undertakings, such as taking part in group activities in music, (2) awareness of social units in the case of a number of children associated for a single purpose, (3) willingness to do one's part in the unison expression of common emotions, (4) ability to enter whole-heartedly and sincerely into



joint enterprises intended for the good of all members, (5) a feeling of common understanding and congeniality when a number of children are united in shared endeavor in pursuit of a goal which all have accepted, and (6) the inclination to subordinate one's self as an individual and to accept the role of follower when that contributes most to the welfare of the greatest number of children.

In addition to the above general goals these same authors listed eighty "subsidiary goals which are contributory to the major goals and which may serve as guides to the teacher in the attainment of the major goals." These goals must not "be considered as a course of study to be followed." They "are constituent elements of the larger objectives. A listing of them should be a great aid to the classroom teacher and the college student who is preparing to teach music or to be a supervisor in that field."<sup>4</sup>

## 2 MEASUREMENT OF MUSICAL TALENT

### Measurement of basic musical talent

Tests of musical aptitude are designed to measure those largely innate musical capacities that constitute the individual's musical inheritance. Aside from the sheer physical endowment that certain types of musical expression demand, there are certain more or less psychological factors that determine an individual's musical talent. The identification of these factors calls for an unusually critical analysis. Without doubt one of the most extensive research programs ever undertaken for the purpose of isolating the elements of native capacity in a special field was that undertaken by Seashore and his students at the State University of Iowa.

The *Seashore Measures of Musical Talent* are a battery of six tests on records for phonographic reproduction designed to measure six elemental abilities upon which response to musical training appears to depend. In the original form of the tests the six elemental abilities measured were (1) sense of pitch, (2) sense of intensity, (3) sense of time, (4) sense of consonance, (5) tonal memory, and (6) sense of rhythm. The testing situation involves careful and critical listening. In each of the tests the individual listens to two tones and is

<sup>4</sup> *Ibid.* p. 116.

then asked to record his judgment concerning the differences that he hears.

On the basis of many constructive criticisms based on the extensive research of his students and published comments concerning the tests, the original Seashore tests were completely rebuilt, appearing in revised form in 1940. In the revision, a test of the individual's ability to distinguish differences in timbre was added. This test involves fifty pairs of tones that differ in their harmonic structure. The time test and the rhythm test were also revised and improved by the use of pure tones of varied duration in the former and by tonal pulses as a means of creating the rhythmic patterns in the latter. The test on consonance was eliminated. In the 1940 revision the tests are arranged in two series: Series A is for general classroom use; Series B is adapted for use with specialized groups and in research studies.

Experience with these tests indicates that the earliest age at which such group measures can be used effectively is at the fifth-grade level. The tests may be administered to groups, the size of the group depending somewhat on the acoustical qualities of the room. Naturally the stimulus must be heard clearly at all times.

According to the author's own statement,

These measures present the following characteristics: they are based on a scientific analysis of musical appreciation and performance; they deal with elements which function in all music; they are standardized for content so that alternate or new series are not needed; they give quantitative results which may be verified to a high degree of certainty; they are economical in that expensive instruments are replaced by phonograph records; they may be used with any language and at any racial or cultural level; they are simple and as nearly self-operative as possible; they are designed for group measurements; they are interpreted in terms of established norms.<sup>5</sup>

The *Kwalwasser-Dykema Music Tests* are quite similar in form and function to the *Seashore Measures of Musical Talent*. The ten tests, designed for use in grades four to twelve, require five phonograph records. The elements measured by the alternate-response technique are: (1) tonal memory, (2) quality discrimination, (3) intensity discrimination, (4) tonal memory, (5) tone discrimination,

<sup>5</sup> Carl E. Seashore, Don Lewis, and Joseph Saetveit, *Manual of Instructions and Interpretations for the Seashore Measures of Musical Talent*. RCA Manufacturing Company, Inc., Educational Department, Camden, N. J., 1939.



(6) rhythm discrimination, (7) pitch discrimination, (8) melodic taste, (9) pitch imagery, and (10) rhythm imagery.

## Measurement of musical memory

Quite in contrast with the two tests of musical talent discussed above is the *Drake Musical Memory Test*, which measures musical aptitude by an entirely different technique. The test is designed for persons of any age above seven whether or not they have had musical

### Excerpts from Score Sheet, Drake Musical Memory Test<sup>6</sup>

- There are 12 trials of entirely different melodies.
- Listen carefully to the first melody in each trial and remember it.
- Listen to what is played next and compare it to the first melody to determine:
  - if it is exactly the same as the first melody,..... if so record S.
  - if it is the same melody played in a different key,..... if so record K.
  - if the time has been changed,..... if so record T.
  - if any notes have been changed,..... if so record N.

S=exactly the SAME melody.  
K=change of KEY.

T=change of TIME.  
N=change of one or more NOTES.

Practice exercise No. 1.

Practice exercise No. 2.

- Record your answers in the score form given below.
- Each trial will be announced by number. When you hear a number announced you will know that a new melody is to be played to which all melodies that follow, in that trial, are to be compared.
- Record your answer during the short pause between each melody. Just time enough will be given to write your answer.
- There is never more than one kind of change in any one comparison.
- Fill in every square. Make the best judgment you can for each comparison.
- Write clearly with capital letters.
- In each trial, listen to the first melody. Wait until more is played and record whether it is the same, or if a change has been made in time, key, or notes.

IF THERE IS ANYTHING YOU DO NOT UNDERSTAND ASK ABOUT IT NOW.

Remember— S=SAME K=KEY change T=TIME change N=NOTE change	1.																		
	2.																		
	3.																		
	4.																		
	5.																		
	6.																		
	7.																		
	8.																		
	9.																		
	10.																		
	11.																		
	12.																		

TOTAL  
ERRORS=FINAL SCORE

SCORED BY \_\_\_\_\_

<sup>6</sup> Raleigh M. Drake, *Musical Memory Test: Score Sheet*. Published by Public School Publishing Co., 1934.

training. The subject listens to twelve melodies played in their proper form or with variations in key, time, or notes. He records his responses to each of the 54 trials on a special record sheet to show whether he recognizes the nature of the difference, if any, between the melody itself, which is played first, and the various versions of it which follow. The above reproduction of the directions and of the response section of the score sheet shows the manner in which the test is given and the manner of recording responses.

### 3 MEASUREMENT OF MUSICAL ACHIEVEMENT

The knowledge, skill, and appreciative outcomes of music instruction are measured by a variety of tests of the pencil-and-paper variety. The majority of these instruments appear to measure the knowledge and skill objectives quite adequately, but they largely neglect the appreciative outcomes. This is not surprising, because of the fact that appreciations are almost impossible to define and extremely difficult to measure.

The *Beach Standardized Music Tests* were among the real pioneers in the measurement of musical achievement. Many of the elements measured by these tests are recognized among the tests of more recent development. The following qualities are scheduled for measurement by the test:

1. Knowledge of essential facts of musical notation.
2. Ability to hear and distinguish the component parts of music, namely the elements of time and tune both in isolated form and in melodies.
3. Aural recognition of the structural elements of music fundamentally necessary for intelligent appreciation.
4. Pitch discrimination.
5. Musical memory.
6. Sight singing through indirect methods.
7. The writing of music.

#### Measurement of musical knowledge

Tests of musical knowledge are variously concerned with musical symbols and terms, time and key signatures, note and rest values, syllables, instrumentation of the orchestra, musical form, and the history and biography of music. Samples are given below to illustrate the measurement techniques rather commonly used. Multiple-choice and simple recall items and matching exercises appear to be most



common among the testing techniques used, although the true-false item is used occasionally. The following samples are somewhat representative of the content of various tests.

### Sample A.<sup>7</sup>

#### COMPOSERS OF FAMOUS COMPOSITIONS




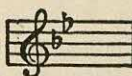
*Directions:* Below are the names of famous compositions. On the lines at the right you are to write the *name of the composer* of each. The sample is marked as it should be.

*Sample:* The Elijah . . . . . Mendelssohn . . . . .

- 
1. March Slav . . . . .
  2. To a Wild Rose . . . . .
  3. The Unfinished Symphony . . . . .
  4. Liebestraum . . . . .

### Sample B.<sup>8</sup>

In the major key signatures below determine what the name of each one is. Find that name above, take its number, and write it in the blank at right of each one, as shown at a. Ready! Go!

<p>a.  <span style="margin-left: 10px;">No. 6 a.</span></p> <p>b.  <span style="margin-left: 10px;">b.</span></p>	<p>g.  <span style="margin-left: 10px;">No. g.</span></p> <p>h.  <span style="margin-left: 10px;">h.</span></p>
---	---

### Sample C.<sup>9</sup>


1. ( ) The viola is an alto horn.
2. ( ) Violins are frequently employed in brass bands.
3. ( ) The first violin section is seated to the left of the conductor.
4. ( ) The harpsichord is one of the predecessors of the piano.

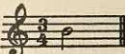
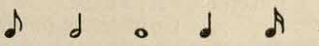
<sup>7</sup>Jacob Kwalwasser, *Kwalwasser Test of Music Information and Appreciation*. Published by Bureau of Educational Research and Service, University of Iowa, 1927.

<sup>8</sup> Clara J. McCauley, *McCauley Examination in Public School Music*. Published by Jos. E. Avent, 1933.

<sup>9</sup> Kwalwasser, *op. cit.*

Sample D.<sup>10</sup>

1  The time signature is  $\frac{2}{4}$   $\frac{3}{4}$   $\frac{4}{4}$   $\frac{3}{8}$   $\frac{9}{8}$  1

1  The note needed is  1

Sample E.<sup>11</sup>

## TEST 1. The Way Musical Instruments Are Played.

**Directions:** Below is a list of musical instruments. They are not all played in the same way. In the blank space opposite each instrument write the letter

A—for each one played by blowing.

B—for each one played by plucking, or picking.

C—for each one played with a bow.

D—for each one played by striking, or shaking.

The sample is marked as it should be.

**Sample:** Ukulele B (because it is played by plucking.)

**Begin here.**

- |                |                  |                   |                      |
|----------------|------------------|-------------------|----------------------|
| 1. Cornet_____ | 7. Trumpet_____  | 13. Trombone_____ | 19. E-flat Alto_____ |
| 2. Banjo_____  | 8. Mandolin_____ | 14. Guitar_____   | 20. Harp_____        |

## Measurement of musical skills

Among the musical skills most commonly measured by various tests are detection of pitch and time errors and recognition of melodies. Illustrations of each are given below. The first represents a type of matching situation and the second a recognition form of item measuring ability to detect errors.

Sample F.<sup>12</sup>

- (e) Below are printed the opening strains of five familiar melodies. After reading or humming them one by one, select the title of each from the list of answers below. Then place the corresponding number in the square at the right of each melody. The sample is correct.

<sup>10</sup> Jacob Kwalwasser and G. M. Ruch, *Kwalwasser-Ruch Test of Musical Accomplishment*. Published by Bureau of Educational Research and Service, University of Iowa, 1924.

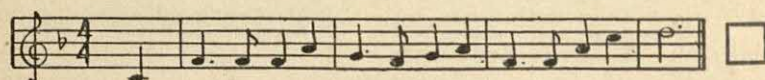
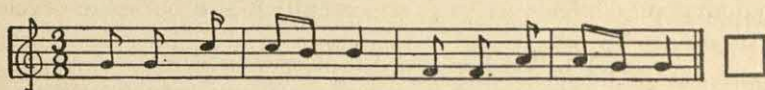
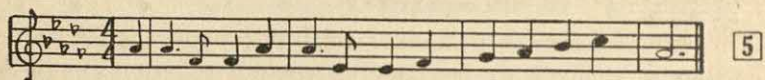
<sup>11</sup> Glenn Gildersleeve and Wayne Soper, *Musical Achievement Test*. Published by Bureau of Publications, Teachers College, Columbia University, 1933.

<sup>12</sup> Frank A. Beach, *Beach Music Test*, Revised. Published by Bureau of Educational Measurements, Kansas State Teachers College, Emporia, 1938.



## LIST OF ANSWERS

- |                    |                          |
|--------------------|--------------------------|
| 1. Silent Night    | 5. America the Beautiful |
| 2. Old Black Joe   | 6. Auld Lang Syne        |
| 3. Santa Lucia     | 7. Star Spangled Banner  |
| 4. Home Sweet Home | 8. America               |



Sample G.<sup>13</sup>

### TEST 3. DETECTION OF PITCH ERRORS IN A FAMILIAR MELODY

DIRECTIONS: The song "America" is written below. One measure has been crossed out because the melody is wrong. Five other measures are wrong. Hum over the melody to yourself and *cross out all five wrong measures*.

Begin here:



### Measurement of music appreciation

Only one test purporting to measure music appreciation, the *Kwalwasser Test of Music Information and Appreciation*, is known to the authors. Its approach is mainly through the testing of knowledges, many of which unquestionably carry appreciative values with them. However, many critics are not convinced that appreciations are measured directly, if, indeed, they can be measured in that manner. In view of the modern emphasis upon music appreciation

<sup>13</sup> Kwalwasser and Ruch, *op. cit.*

for all pupils, it is unfortunate that the appreciative types of outcomes are not subject to satisfactory measurement.

## 4 CHARACTERISTICS AND AIMS OF ART EDUCATION

### General trends in art education

The theory of art for art's sake, which dominated the field of art education for many years, has largely given way of late to the theory that all pupils should receive an opportunity in art courses to develop a sensitivity to beauty and critical taste in evaluating art objects. Hence, art is no longer thought of as a field only for the talented few. Creative self-expression, especially in the lower grades, and correlation of art with other activities of the school are important modern trends. These trends involve the use of a wide variety of art materials in the classroom. Extension of the content of art education courses beyond the drawing and painting, which largely constituted the curriculum in the past, particularly to industrial arts is another trend worthy of note. Last, and perhaps most important, the appreciative aims of art education have increasingly come to the front.<sup>14</sup>

Modern aims and purposes of art education in relation to current social needs are summarized effectively in the following statement: "Art in the modern school should aim both to stimulate in the child the experience of creating and to help him to improve the manner in which he expresses himself through creative processes; at the same time, it should aim to stimulate in him the experience of appreciating by acquainting him systematically with fine examples of the arts of various peoples, both of the present and of the past."<sup>15</sup>

### General outcomes of art education

Three general outcomes of art education appear to be of major importance: (1) information, (2) appreciation, and (3) expression.<sup>16</sup> It is quite probable that art appreciation is not necessarily taught, although real appreciation may be considered to rest to a large

<sup>14</sup> Robert S. Hilpert, "Changing Emphases in School Art Programs," *Art in American Life and Education*, Fortieth Yearbook of the National Society for the Study of Education. Public School Publishing Co., Bloomington, Ill., 1941. p. 452-53.

<sup>15</sup> Leon L. Winslow, *The Integrated School Art Program*. McGraw-Hill Book Co., Inc., New York, 1949. p. 43.

<sup>16</sup> Walter H. Klar, Leon L. Winslow, and C. Valentine Kirby, *Art Education in Principle and Practice*. Milton Bradley Co., Springfield, Mass., 1933.



degree upon the broader aspects of information. There will still remain something in the truly artistic product which sheer information does not entirely explain. The third major objective might be better expressed as exploration. Not many potentially great artists are discovered in the elementary-school classroom, but practically all the great artists there are have come up through this avenue. Not everyone can express himself effectively in artistic form, but everyone has a right to explore for himself the fields of human expression in the hope that his own hidden talent may be uncovered. Art talent and achievement tests have distinct contributions to make in this field.

### Specific outcomes of art education

The elementary-school art course faces the responsibility for bringing to the child a fourfold artistic experience. These categories of experience were suggested by Kirby<sup>17</sup> in a discussion of aims and tendencies in art education. The first is the *graphic* experience, which expresses itself in representative drawing, illustrative and imaginative drawing, nature drawing, and other related forms. The second is the *thoughtful* experience, involving the constructive, decorative phases of artistic expression. The third experience involves the *acquisition of motor skill* in expression. The fourth is the *emotional* experience, which involves the appreciation of the arts.

A somewhat more detailed expression of outcomes of instruction in art is given in the accompanying outline adapted from a course of study covering the first six years of the elementary school. It will be noted that this course is organized around four groups of outcomes which are quite similar to the artistic experiences presented in the previous paragraph.

#### OUTCOMES OF ART INSTRUCTION<sup>18</sup>

##### A. Fruitful knowledge

Functional information

Practical relation of art to everyday life (clothing, home, town, or city, etc.)

<sup>17</sup> C. Valentine Kirby, "Aims and Tendencies." *Pennsylvania School Journal*, 77:501-2; April 1929.

<sup>18</sup> Adapted from W. G. Whitford, *An Introduction to Art Education*. Century Co., New York, 1929.

Understanding of elements and principles of art and their adaptation to everyday use

Knowledge of construction and industrial processes involved in art training

Acquaintance with art of other countries

**B. Attitudes, interests, and appreciations**

Civic consciousness (civic pride)

Appreciation and understanding of beauty in modern products of all kinds

Interest in art museums, travel, and further study

Interest in the civic, domestic, and social service of art

**C. Mental technique**

Good taste, discriminating judgment, ability to select and choose wisely

Creative ability, originality, initiative, imagination, keen observation

Ability to analyze works of art and to understand the factors of beauty in production

Keener observation; beauty of nature and fine things of art

**D. Right habits and skills**

Constructive thinking and planning

Systematic organization

Practical technique

Coordination of mind, hand, and eye

Freedom and spontaneity

Order, neatness

Body and mind training

Self-activity

Worthy use of leisure time

## **5 MEASUREMENT OF ART ABILITIES AND ACHIEVEMENT**

Three types of tests may be distinguished in the field of art education: (1) drawing scales and tests, (2) art appreciation tests, and (3) art abilities tests. As many of the art tests cannot be illustrated easily, brief descriptions of a few representative scales and tests and one illustration are given on the following pages to familiarize the student with measurement devices in this field.



## Drawing scales and tests

Several rating scales for use in the evaluation of art achievement are now available. Such scales must depend, as do all scales, upon the representative nature of the specimens selected for presentation and the skill of judges in using the scales. Evidence from a study of the values of drawing scales indicates that their use reduces the inaccuracy of ratings to about one-half of that obtained when no scale is used.<sup>19</sup>

The *Kline-Carey Drawing Scales* consist of series of samples for measuring (1) representation, and (2) design and composition. The first series uses such subject matter as a house, a running boy, a tree in silhouette, and a rabbit in scales having 14 samples, while the second uses the themes of illustrations, posters, structural designs, and borders.

## Tests of art appreciation

The increasing stress placed upon the appreciative outcomes of art instruction of recent years results in a significant place for art appreciation tests among evaluative tools. Two tests of art judgment or talent are worthy of brief discussion here—the *Meier Art Judgment Test* and the *McAdory Art Test*.

In the *Meier Art Judgment Test*, which may be given as an individual or as a group test, the pupil is confronted with 100 pairs of artistic specimens adapted from many sources. One of each pair of specimens has been changed in some specific element from the original form. The exact feature changed is specified in the record sheet on which the pupil records his reactions. A consideration of the complete series of paired specimens insures a comprehensive sampling of the various elements that enter into æsthetic judgment. According to the evidence obtained by the author, this test measures the sensitivity of the individual to the effect that the composition as a whole produces on the observer. In order to give a better idea of the exact nature of specimens and the accompanying record sheets, a single pair of the etchings is reproduced here along with a brief sampling of seven items from the Test Record Sheet. The pair of specimens reproduced here is used with item 49 in the record sheet. In this item,

<sup>19</sup> Fowler D. Brooks, "The Relative Accuracy of Ratings Assigned with and without the Use of Drawing Scales." *School and Society*, 27:518-20; April 28, 1928.

the presence or absence of horns is the point for special consideration in making the judgment. The scoring key lists the drawing with horns as the one of greater merit.

### Excerpts from Meier Art Tests, I, Art Judgment <sup>20</sup>

#### DIRECTIONS

In the accompanying booklet are pictures arranged in pairs, the two in each pair being very nearly alike. They differ only in *one* respect and you are told *what* that is in each case on pages 1, 2, and 3 of this blank.

You are to compare the two pictures in each pair, noting the unlike portion, and then decide which one is *better* (more pleasing, more artistic, more satisfying). Do not hurry. Study each pair carefully in turn.

Indicate your preference by making an X in the circle under *Left*, if you decide that the left-hand picture is *better*, or in the circle under *Right* if you believe that the right hand one is more desirable.

Examples of proper marking: (pictures not illustrated).

Left Right No.

☒ ☐ A Presence or absence of tree. (This would mean that you prefer the left-hand picture)

☐ ☒ B Treatment of waves. (This would mean that you prefer the right-hand picture)

Select the better one in every pair. Do not omit any. If unable to decide within a reasonable time mark the place and return to that one later.

Left	Right	Pair No.	Difference
<input type="radio"/>	<input type="radio"/>	1	Arrangement of wall and foreground
<input type="radio"/>	<input type="radio"/>	2	Foreground
<input type="radio"/>	<input type="radio"/>	49	Inclusion or omission of the horns
<input type="radio"/>	<input type="radio"/>	50	Arrangement in picture of the woman and umbrella
<input type="radio"/>	<input type="radio"/>	51	Position of the figures
<input type="radio"/>	<input type="radio"/>	99	Direction of pine tree's main branch
<input type="radio"/>	<input type="radio"/>	100	Treatment of the water



49

The *Meier Art Judgment Test*, which supersedes the *Meier-Sea-shore Art Judgment Test*, is made up of a smaller number of carefully selected items. The *McAdory Art Test* is somewhat similar to the *Meier Art Judgment Test*, but has only 72 pairs of plates, 24 of which

<sup>20</sup> Norman C. Meier, *The Meier Art Tests, I, Art Judgment*. Published by Bureau of Educational Research and Service, University of Iowa, 1940.



are in color, and calls for reactions to a wide variety of materials, such as furniture, clothing, and rugs.

### Art abilities tests

Two tests that purport to measure art abilities mainly the outcomes of art instruction are the *Lewerenz Test in Fundamental Abilities of Visual Arts* and the *Knauber Test of Art Ability*. Their major values seem to be at the junior and senior high-school levels, although the first is designed for use as low as the third grade.

The *Lewerenz Test in Fundamental Abilities of Visual Arts* is designed to measure aspects of art ability in nine areas: (1) recognition of proportion, (2) originality of line drawings, (3) observation of light and shade, (4) knowledge of subject-matter vocabulary, (5) visual memory of proportion, (6) analysis of problems in cylindrical perspective, (7) analysis of problems in parallel perspective, (8) analysis of problems in angular perspective, and (9) recognition of color.

### Applied art

Education will have failed in much of its social responsibility if it allows the child to leave the school without developing in him a rather definite love of the fine arts, even though it may be on a relatively low level. Not everyone can or should become a musician, a painter, or a sculptor, but almost everyone has the essential elements which make for a love and appreciation for the beautiful which he himself may be unable to produce. Instruction in the fine arts should cultivate and develop these elements. Furthermore, such instruction has a rather definite responsibility for making the arts function in real life in matters of personal adornment and in the planning and decorating of the home. The general cultural level of the population will be raised as this point is recognized and applied in instruction in the fine arts.

### Topics for Discussion

1. In your opinion is there any reason to assume that achievement in the fine arts may not be objectively measured? Support your answer.
2. Is the measurement of musical talent or artistic talent any more difficult or basic than the measurement of aptitudes in any other complex field?

3. What are the major types of aims in music education?
4. Which of the aptitude tests discussed here seem to be most soundly grounded in critical research?
5. Briefly discuss and illustrate the manner in which musical knowledge and musical skills are measured.
6. What is the status of standardized tests of musical appreciation?
7. What similarities in the basic problems of measurement do you see in the fields of music and art?
8. What are the major classes of general outcomes in art instruction?
9. Which of the art tests described here seem most adequately to measure the major features of accomplishment in art?
10. Distinguish clearly between art appreciation and art abilities tests with respect to their functions and forms.

### Selected References

- BARNES, MELVIN W. "A Technique for Testing Understanding of the Visual Arts." *Educational and Psychological Measurement*, 2:349-52; October 1942.
- BROOKS, B. MARIAN, AND BROWN, HARRY A. *Music Education in the Elementary School*. New York: American Book Co., 1946.
- BROOM, M. E. *Educational Measurements in the Elementary School*. New York: McGraw-Hill Book Co., Inc., 1939. p. 236-44.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 335-48.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p. 143-57.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938. p. 43, 119-21.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. p. 257-64.
- DEAN, CHARLES D. "Predicting Sight-Singing Ability in Teacher-Education." *Journal of Educational Psychology*, 28:601-8; November 1937.
- DUNKEL, HAROLD B. *General Education in the Humanities*. Washington, D. C.: American Council on Education, 1947. Chapter 5; p. 312-21.
- FARNSWORTH, PAUL R. *Musical Taste: Its Measurement and Cultural Nature*. Education-Psychology Series, Vol. II, No. 1. Stanford, Calif.: Stanford University Press, 1950.
- FAULKNER, RAY N. "Evaluation in Art." *Journal of Educational Research*, 35:544-54; March 1942.



- FAULKNER, RAY N. "Standards of Value of Art." *Art in American Life and Education*. Fortieth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1941. Chapter 27.
- GILDERSLEEVE, GLENN. "Standards and the Evaluation and Measurement of Achievement in Music." *Music Education*. Thirty-Fifth Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1936. Chapter 19.
- GRANT, PARKS. *Music for Elementary Teachers*. New York: Appleton-Century-Crofts, Inc., 1951.
- GRAVES, MAITLAND. "What is Your IQ in Design?" *Art Instruction*, 3:11-14; April 1939.
- HARRIS, CHESTER W., BETTELHEIM, BRUNO, AND DIEDERICK, PAUL B. "Aspects of Appreciation." *Appraising and Recording Student Progress*. New York: Harper and Brothers, 1942. p. 276-312.
- HENDRICKSON, GORDON, AND STRATEMEYER, CLARA G. "Music Education." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 761-72.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. p. 288-307.
- KLAR, WALTER H., WINSLOW, LEON L., AND KIRBY, C. VALENTINE. *Art Education in Principle and Practice*. Springfield, Mass.: Milton Bradley Co., 1933.
- KWALWASSER, JACOB. *Tests and Measurements in Music*. Boston: C. C. Birchard and Co., 1927.
- MEEKER, RONALD W. *Musical Aptitudes, Achievement, and Appreciation as Related to Pupil Participation*. Unpublished Master's Thesis. Iowa City: State University of Iowa, 1946.
- MEIER, NORMAN C. "Diagnosis in Art." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1935. Chapter 22.
- MEIER, NORMAN C. "Recent Research in the Psychology of Art." *Art in American Life and Education*. Fortieth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1941. Chapter 26.
- MUNRO, THOMAS, chairman. *Art in American Life and Education*. Fortieth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1941.
- MURSELL, JAMES L. *Music and the Classroom Teacher*. New York: Silver Burdett Co., 1951.
- MURSELL, JAMES L., AND OTHERS. "The Measurement of Understanding in the Fine Arts." *The Measurement of Understanding*. Forty-Fifth

- Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1946. Chapter 10.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. Chapter 8.
- SCHOEN, MAX. "Report of the Committee on Music Tests and Measurements." *Proceedings of the Music Teachers National Association*. Oberlin, Ohio: The Association, 1935. p. 320-50.
- SCHULTZ, HAROLD A., ROOS, FRANK J., AND MOORE, J. E. "Art Education." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 64-72.
- SEASHORE, CARL E. "The Discovery and Guidance of Musical Talent." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1935. Chapter 21.
- SQUIRE, RUSSEL N. *Introduction to Music Education*. New York: Ronald Press Co., 1952.
- STANTON, HAZEL M. *Prognosis of Musical Achievement*. Studies in Psychology, Vol. I, No. 4. Rochester, N. Y.: Eastman School of Music, University of Rochester, 1929.
- TODD, JESSIE M. "A Test in Art for Grade Children." *School Arts Magazine*, 30:365-68; February 1931.
- UHL, WILLIS L. "Contributions of Research to Special Methods: Music and Art." *The Scientific Movement in Education*. Thirty-Seventh Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1938. Chapter 14.
- UHL, WILLIS L., chairman. *Music Education*. Thirty-Fifth Yearbook of the National Society for the Study of Education, Part II. Bloomington, Ill.: Public School Publishing Co., 1936.
- WATKINS, JOHN G. "Objective Measurement of Instrumental Performance." *Teachers College Record*, 44:376-77; February 1943.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapters 17-18.
- WHITFORD, WILLIAM G. *An Introduction to Art Education*. New York: Century Co., 1929.
- WOODS, ROY C., AND MARTIN, LUREATA R. "Testing in Musical Education." *Educational and Psychological Measurement*, 3:29-42; Spring 1943.



## *Measuring and Evaluating in Health and Physical Education*

THIS CHAPTER presents a brief summary of the following aspects of measurement and remediation in health and physical education:

- A. Status and aims of health education.
- B. Measurement and evaluation in health education.
- C. Diagnostic and preventive measures for the improvement of health.
- D. Philosophy and objectives of physical education.
- E. Measurement in physical education.
- F. Diagnosis of physical condition.

The rather closely related fields of health and physical education are exceedingly important in the school, although they seem not to be as much favored in the curricular setup in most schools as are the academic areas. The national and economic importance of good health is obvious, but the loss to society from illness and unnecessary death is not so apparent as the loss to the individual. Physical education perhaps occupies a more important place in the society of today than was true earlier in the history of man because of the modern need for physical activities to counteract the effects of the physically inactive lives led by many persons.

### **1** "SCOPE AND AIMS OF HEALTH EDUCATION

Considered in its broadest sense, health education includes much more than can be treated in this chapter. The mental health, or mental hygiene, aspect is considered in Chapter 11. Some of the

health education activities, such as health service, mental hygiene of the classroom, and recreation, are not measurement problems as much as they are supervisory or administrative problems. This chapter deals primarily with health education and physical education measurement, with some attention to the aims and objectives of these fields. Although they are in the main treated separately here, health and physical education are not equally inclusive terms. Instead, the latter may be considered as one aspect of the former.

## Scope of health education

Strang stated the scope of health education as follows:

Health education is concerned with healthy growth, the prevention of disease, the correction of physical impairments, and the building of a healthful environment. At its best health education builds physical security and contributes to self-realization, social security, and the welfare of society.<sup>1</sup>

This statement makes clear the great responsibility of the school for the health education of the child, especially because the child not only undergoes physical as well as intellectual experiences of rather broad scope in his school activities but also because his school actions and personality are influenced by his physical activities outside of and beyond the control of the school.

## Aims of health education

The aims stated by a joint committee of educators and physicians<sup>2</sup> indicate the purposes underlying health education.

1. To instruct children and youth so that they may conserve and improve their own health.
2. To establish in them the habits and principles of living which throughout their school life and in later years will aid in providing that abundant vigor and vitality which are a foundation for the greatest possible happiness and service in personal, family, and community life.

<sup>1</sup> Ruth Strang, "Health Education." *Encyclopedia of Educational Research*, Revised edition. Macmillan Co., New York, 1950. p. 529.

<sup>2</sup> *Health Education*. Report of Joint Committee on Health Problems in Education of the National Education Association and the American Medical Association, Second revision. National Education Association, Washington, D. C., 1941. p. 15.



3. To promote satisfactory habits and attitudes among parents and adults thru parent and adult education and thru the health education program for children, so that the school may become an effective agency for the advancement of the social aspects of health education in the family and in the community as well as in the school itself.
4. To improve the individual and community life of the future; to insure a better second generation, and a still better third generation; to build a healthier and fitter nation and race.

Comparatively minor differences in topics comprising the health curriculum are found from school to school in the elementary grades. The major areas of study have to do with: (1) such health habits as cleanliness, food and nutrition, sleep and rest, posture and exercise, dental hygiene, ventilation; clothing; first aid and safety; and effects of alcohol and narcotics, and (2) attitudes of courage, helpfulness, consideration of others, independence, adaptability, and enjoyment of daily living.<sup>3</sup>

## 2 MEASUREMENT AND EVALUATION IN HEALTH EDUCATION

Classroom testing in the field of health education by the use of paper-and-pencil tests has not attained any significant state of development. Since standardized health tests are not numerous, the discussion to follow is necessarily brief. Some of the measuring instruments discussed in the latter part of this chapter under the heading of physical education have at least indirect significance as health measures.

### Health knowledge tests

Several good health knowledge tests have been published since 1930, but most of the older tests doubtless have little significance today because of the tremendous advances that have been made in nutrition since their publication and the importance of correct knowledge concerning nutrition as a basis for making dietary decisions.

"Strang pointed out that norms are of no great significance for health knowledge tests because the discovery of individual pupil

<sup>3</sup> Bernice E. Leary, *A Survey of Courses of Study and Other Curriculum Materials Published since 1934*. U. S. Office of Education Bulletin, 1937, No. 31. U. S. Government Printing Office, Washington, D. C., 1937.

variations and their meaning is much more fundamentally important than are comparisons of individual or group test performance with a norm.<sup>4</sup>

Sample items from two health knowledge tests are given in accompanying illustrations as representative of the testing technique and content of modern tests in this field. One of the best-known tests of this type is the *Gates-Strang Health Knowledge Test*, which is a revision of an earlier test by the same authors.

### Excerpts from Gates-Strang Health Knowledge Test<sup>5</sup>

- |  |  |
|--|--|
| <p>1. Of these five foods, the most important one for children is</p> <p>a. Meat. .... a</p> <p>b. Butter. .... b</p> <p>c. Fish. .... c</p> <p>d. Sugar. .... d</p> <p>e. Milk. .... e</p>  | <p>40. You use up the most calories when you are</p> <p>a. Asleep in bed. .... a</p> <p>b. Lying in bed awake. .... b</p> <p>c. Sitting still. .... c</p> <p>d. Standing. .... d</p> <p>e. Running .... e</p>  |
| <p>12. Automobile accidents and hard falls are not likely to happen to you when you</p> <p>a. Are in a great hurry. .... a</p> <p>b. Take dares. .... b</p> <p>c. Are very tired and sleepy. .... c</p> <p>d. Are worried about things at home or school. .... d</p> <p>e. Look where you are going and keep your wits about you. .... e</p> | <p>50. When a person is in good health, his heart</p> <p>a. Beats fast or slowly according to the needs of the body as a whole. .... a</p> <p>b. Always beats at the same rate ... b</p> <p>c. Never skips a beat. .... c</p> <p>d. Always sends the same amount of blood to each part of the body. .... d</p> <p>e. Beats more slowly when a person walks or runs than when he sits still. .... e</p> |

<sup>4</sup> Strang, *op. cit.* p. 536.

<sup>5</sup> Arthur I. Gates and Ruth Strang, *Gates-Strang Health Knowledge Test*, Grades 3 to 8. Published by Bureau of Publications, Teachers College, Columbia University, 1937.



Excerpt from Progressive Tests in Social and Related Sciences <sup>6</sup>

26. If the water of a pool or stream looks clear, it is safe to drink. T F<sup>26</sup>
27. A person who works and plays in an overheated room is not likely to have any colds. T F<sup>27</sup>
28. People who drink a great deal of alcohol do not live as long, on the average, as do other people. T F<sup>28</sup>
29. If you have worked or played until you are very warm, you should cool off quickly. T F<sup>29</sup>

## Health attitudes inventories

Comparatively little effective work has been done in the measurement of health attitudes of pupils, except for the *Health Attitudes Inventory* illustrated on page 286. Health attitudes can be approached from two standpoints: (1) pupil attitudes toward health practices and beliefs in certain courses of action, and (2) pupil likes and dislikes for various types of foods, activities, and health practices. The same weakness is inherent in these instruments as in attitudes scales in general—there is little evidence to support the belief that expressed attitudes necessarily are borne out in terms of conduct.

The *Health Awareness Test*, from which a few sample items are shown below, is one of the fairly recent publications of a closely related type. It is the result of research in the health measurement field carried on by the American Child Health Association.

## Health evaluation inventories

A series of health inventories was recently developed by the Co-operative Study in General Education for use from Grade 9 through the college years. The six inventories are numbered and entitled: I, Health Activities; II, Health Information; III, Health Interests; IV, Health Attitudes; V, Analyzing Health Problems; and VI, Judg-

<sup>6</sup> Georgia S. Adams and John A. Sexson, *Progressive Tests in Social and Related Sciences, Test 5, Health and Safety*, Elementary. Published by California Test Bureau, 1946.

ing Sources of Information in Health Problems. Inventories III and IV are represented by brief excerpts in Chapter 11, pages 288 and 286, and other illustrations from the series appear in Chapter 9, page 224, of this volume. These instruments represent undoubtedly the most extensive attempt to evaluate instructional outcomes in health education. They are intended to depict pupil needs for instruction and guidance in this area. Norms are not provided.

### Excerpts from Health Awareness Test <sup>7</sup>

(Direction for matching test given on blackboard and orally.)

- |   |                  |
|---|------------------|
| 1. Keep from breeding                   | ( ) Wet feet     |
| 2. Should not touch other people's food | ( ) Bad cold     |
| 3. Blow the nose gently, not hard       | ( ) Bedroom      |
| 4. Keep covered                         | ( ) Garbage pail |
| 5. Scald with boiling water             | ( ) Flies        |
| 6. Should be very clean                 | ( ) Sore throat  |
| 7. Should not be too warm               | ( ) Babies' milk |
|   | ( ) Sick people  |
|   | ( ) Whiskey      |
|   | ( ) Dirty dishes |

*DIRECTIONS: If statement number 1 is true, put a circle around the T, but if it is false, put a circle around the F. Do the same for all the statements.*

- 
- |  |   |   |
|--|---|---|
| 1. Candy should be eaten only at the end of a meal. ....   | T | F |
| 2. An orange, a glass of milk, and hot cooked whole wheat cereal is a better breakfast than an orange, a glass of milk, and puffed wheat. .... | T | F |
| 3. Hot cinnamon rolls or white rolls, fresh and hot, are the best kind of bread for boys and girls. ....                                       | T | F |

### Physical examinations

Physical examinations in connection with the evaluation of pupil health will be mentioned only briefly here, for they obviously are not

<sup>7</sup> Raymond Franzen, Mayhew Derryberry, and W. A. McCall, *Health Awareness Test*. Published by Bureau of Publications, Teachers College, Columbia University, 1933.



the province of the classroom teacher. Health defects often come to light in these examinations, although the annual physical check-ups in some schools may be so perfunctory as to overlook serious health defects.

### 3 PREVENTION AND DIAGNOSIS IN HEALTH EDUCATION

Diagnosis in health education perhaps more than in any other instructional field must be considered both from the standpoint of school diagnosis and from the usual standpoint of individual pupil diagnosis. Class diagnosis has been discussed in several chapters of this volume, but school diagnosis is here mentioned for the first time.

Diagnosis in health education perhaps more than in any other considered as related to health is a problem of such constant significance that objective tests can be expected to contribute only in a relatively indirect manner. They cannot be of value in diagnosing contagious diseases and other health conditions that demand immediate attention when cases are found. The diagnostic significance of results from health knowledge tests and health attitudes inventories is not specific. The former are survey rather than specifically diagnostic tools and the latter suffer from the fact that their results are not necessarily indicative of health behavior. Therefore, the major diagnostic possibilities in the field of health education are probably to be found elsewhere than in the standardized test, although the diagnostic significance of the *Health Inventories* discussed briefly above appears to be considerable.

The physical examination has diagnostic significance, of course, but the advantages of continuous measurement and diagnosis are lost when such examinations are made only at infrequent intervals. It is possible, however, for the teacher to supplement the physical examination through his opportunities for the daily observation of pupils in the school. The teacher's place as a diagnostician, non-technical though his diagnoses may be, is fundamentally important. The teacher should recognize that sore throat, vomiting, skin rashes, and various evidences of contagious colds frequently indicate the desirability of an immediate dispatch of the pupil to the school nurse or to his home. It is through the abilities of teachers to

diagnose illness, although not necessarily its specific nature, that individual and group pupil health are protected. The opportunity for such diagnoses of pupil health is usually provided by the morning inspection.

Diagnosis by the teacher can also be made for less immediately important health conditions as the result of continuous observation of pupils. For example, visual defects may be recognized through postural conditions during reading; auditory defects are sometimes major causes of poor spelling; malnutrition frequently results in physical abnormality; goiter of some types is evidenced in a swelling of the thyroid gland in the throat; and such nervous ailments as epilepsy and chorea furnish unmistakable signs. The teacher can often supplement the work of the health agencies of the school by constant alertness for such signs of health defects and by consulting with qualified authorities or referring the case to the proper agencies when he recognizes characteristics he believes to be symptomatic of defects needing remediation.

Prevention as a phase of diagnosis is clearly important in health education. Isolation of pupils with contagious diseases, periodic chest X-rays, and immunization of pupils against smallpox, diphtheria, and typhoid fever are preventive responsibilities now accepted by the schools in many communities. Provision of healthful school conditions and a desirable type of school atmosphere and morale are also important as preventive measures.

#### 4 OBJECTIVES OF PHYSICAL EDUCATION

Physical education during the past two decades has come to be thought of as making an increasingly valuable contribution to the educational process, and its philosophy has consequently been dominated recently by broader aims than were generally held previously. The colleges and secondary schools have better-organized programs than do the elementary schools, as less attention has been devoted to physical education for elementary-school children than for high-school and college students. The statement of aims that follows represents the modern philosophy concerning the contribution physical education should make to the attainment of desirable educational outcomes in the pupil.



The general objectives of physical education listed by LaPorte<sup>8</sup> indicate the types of pupil outcomes to which a good physical education program should lead.

1. The development of fundamental skills in aquatic, gymnastic, rhythmic, and athletic activities for immediate educational purposes—physical, mental, and social.
2. The development of useful and desirable skills in activities suitable as vocational interests for use during leisure time.
3. The development of essential safety skills and the ability to handle the body skillfully in a variety of situations for the protection of self and of others.
4. The development of a comprehensive knowledge of rules, techniques and strategies in the above activities suitably adapted to various age levels.
5. The development of acceptable social standards, appreciations and attitudes as the result of intensive participation in these activities in a good environment and under capable and inspired leadership.
6. The development of powers of observation, analysis, judgment, and decision through the medium of complex physical situations.
7. The development of the power of self-expression and reasonable self-confidence (physical and mental poise); by mastery of difficult physical-mental-social problems in supervised activities.
8. The development of leadership capacity by having each student within the limits of his ability, assume actual responsibility for certain activities under careful supervision.
9. The elimination of remediable defects and the improvement of postural mechanics insofar as these can be influenced by muscular activities and health advice, based on adequate physical and health diagnosis.
10. The development of essential health habits, health knowledge and health attitudes as the result of specific instruction in health principles and careful supervision of health situations.

## 5 MEASUREMENT IN PHYSICAL EDUCATION

Persons interested in testing and evaluating various aspects of physical ability and skill will find almost no commercially published paper-and-pencil tests for those purposes. On the other hand, they will find voluminous reports of testing and evaluative techniques in

<sup>8</sup> William R. LaPorte, "The Ten Major Objectives of Health and Physical Education." *California Physical Education, Health and Recreation Journal*, January 1936. p. 6.

the educational literature, both books and journals. It should be apparent that measures of physical fitness, motor ability, and physical skills must be conducted by means of physical and medical measurements and tests and by observation of behavior in situations involving physical activity rather than by the use of standardized tests.

### Tests of general physical qualities

Tests of such qualities, commonly thought to be inherited, as motor ability, physical capacity, and athletic ability are considered under this heading. Each of these tests consists in the main of various measures of motor abilities and physical skills combined into a composite score. Their results are useful variously as a basis for classifying pupils into groups for physical education and for predicting levels of physical attainment.

Rogers devised a series of physical tests for administration to individual pupils from which two types of derived scores are obtained—a strength index and a personal fitness index.<sup>9</sup> The tests, having different procedures in some instances for boys and girls, are accompanied by tables of normal strength indices differentiated for age and sex groups. Their major purpose is to determine deficiencies and to facilitate classification of pupils into groups having common remedial needs. No presentation of these tests can be given here because of their detailed nature.

A test of general motor capacity based on various specific tests of track and field events and of strength was developed by McCloy.<sup>10</sup> Something approaching a profile of the individual's general capacity is furnished by the results of this test, which has particular value in the prediction of ultimate levels of attainment.

The *Iowa Revision of the Brace Test of Motor Ability* consists of 21 physical stunts yielding a composite score of motor educability.<sup>11</sup> The tests are so devised and set up that pupils can do the scoring and the recommended procedure is that one half of the class score the

<sup>9</sup> Frederick R. Rogers, *Physical Capacity Tests*. A. S. Barnes and Co., New York, 1938.

<sup>10</sup> C. H. McCloy, "The Measurement of General Motor Capacity and General Motor Ability." *Research Quarterly of the American Physical Education Association*, 5:46-61; March 1934.

<sup>11</sup> C. H. McCloy, "An Analytical Study of the Stunt Type Test as a Measure of Motor Educability." *Research Quarterly of the American Physical Education Association*, 8:46-55; October 1937.



other half on performance of the stunts and that the two groups then be reversed. Scores are interpreted in terms of *T*-score values of the type dealt with in Chapter 13.

Johnson devised a test of motor educability that has values for the sectioning of classes.<sup>12</sup> Although it is slower and more difficult to administer than the *Iowa-Brace Test* it is rated as probably the best available test of motor educability.

## Cardiovascular tests

Good and poor physical condition can be determined by cardiovascular tests involving pulse rate and blood pressure under varying conditions of rest and fatigue. Several of the tests of this type based on pulse counts are sufficiently easy to administer and require so little equipment that they are subject to use by the skilled teacher.<sup>13</sup> The significance of such tests is somewhat reduced by the fact that most of them measure only one type of physiological efficiency, whereas some other of the important variables of blood pressure, pulse rate, and related functions are not well enough understood at this time to be included in these tests.

## Posture tests

Posture tests cannot be administered in a routine manner in the usual school situation because of their complex nature. Most of the tests of posture are based on comparisons of pupil silhouettes with silhouettes representing standard posture or representing several degrees of postural merit from very poor to excellent. Because of the wide variability in posture and the incomplete nature of evidence on the question, Ashbrook and his colleagues suggested that teachers should probably not attempt to make pupils conform to any pattern considered desirable.<sup>14</sup>

<sup>12</sup> Granville B. Johnson, "Physical Skill Tests for Sectioning Classes into Homogeneous Units." *Research Quarterly of the American Physical Education Association*, 3:128-36; March 1932.

<sup>13</sup> C. H. McCloy, *Tests and Measurements in Health and Physical Education*. F. S. Crofts and Co., New York, 1939. Chapter 20.

<sup>14</sup> Willard P. Ashbrook, Anna Espenschade, and Frederick W. Cozens, "Physical Education—Measurement." *Encyclopedia of Educational Research*, Revised edition. Macmillan Co., New York, 1950. p. 836.

## General achievement scales

General achievement scales have been developed for the measurement of ability in various sports activities. Their purposes are to stimulate pupil interest and performance, to determine the sports skills of individual pupils and groups, and to diagnose deficiencies. Such scales are highly time-consuming, however, and have not been adequately validated.<sup>15</sup>

## Knowledge and information tests

Paper-and-pencil tests of knowledge and information in specific sports activities and comprehensively for all activities have appeared in the physical education journals but have not been published commercially in standardized form. The following illustrations indicate the manner in which various objective item forms are adaptable to measurement of knowledge and information in this field.<sup>16</sup>

### TRUE-FALSE ITEMS

Encircle the correct answer:

- T F The follow-through in a golf drive determines the accuracy of the flight of the ball.
- T ? F There are African negro tribes who have athletes able to high jump to heights greater than the present American record.

Encircle the correct answer. If the answer is false, cross out the word which makes it false and insert the word that makes it true.

anopheles

- T F The ~~culex~~ mosquito is the transmitter of the malaria germ.

Underline T if the statement is true and F if the statement is false. If the converse of the statement is true, underline CT; if the converse is false, underline CF.

- T F CT CF Low arches are always painful.

<sup>15</sup> Inasmuch as these scales are too highly specialized to warrant presentation or illustration here, the student should refer to the bibliography at the end of this chapter for source materials.

<sup>16</sup> McCloy, *op. cit.* p. 190-97.



## MULTIPLE-CHOICE ITEMS

Place an X in the space before the phrase which correctly completes the statement.

The world's record for the mile run is approximately:

- \_\_\_\_\_ 3' 20"
- \_\_\_\_\_ 4' 6"
- \_\_\_\_\_ 8' 11"
- \_\_\_\_\_ 2' 19"

Check the one or more correct answers under each statement:

According to the currently accepted "best" form for the shot-put (for a right-handed putter):

- \_\_\_\_\_ In the hop, the right foot alights well before the left foot.
- \_\_\_\_\_ The shot should remain as close to the neck as possible.
- \_\_\_\_\_ The reverse is of no importance, and is just a traditional movement.
- \_\_\_\_\_ The shot should be held deep in the palm of the right hand.
- \_\_\_\_\_ The best angle (to the ground) of the putting effort is approximately forty-one degrees.

## MATCHING EXERCISES

In the following questions, write the number belonging to the approximately correct date in the first space and the number corresponding to the correct name in the second space.

At about the year:

- \_\_\_\_\_, a physical education program was introduced at the Philanthropium in Dessau by \_\_\_\_\_.
- \_\_\_\_\_, physical education was established at the Round Hill School in the United States by \_\_\_\_\_.
- \_\_\_\_\_, a department of physical education was opened in the Y. M. C. A. Training School at Springfield, Massachusetts, under the guidance of \_\_\_\_\_.
- \_\_\_\_\_, the King of Denmark appointed as professor of physical education in the university \_\_\_\_\_.
- \_\_\_\_\_, the modern Olympic Games were revived, largely because of the work of \_\_\_\_\_.

Dates		Names	
1. 1776	6. 1887	1. Basedow	6. Hitchcock
2. 1799	7. 1897	2. Beck	7. Jahn
3. 1804	8. 1902	3. Bukh	8. Ling
4. 1810	9. 1906	4. de Coubertin	9. McCurdy
5. 1823	10. 1924	5. Gulick	10. Nachteggall

### COMPLETION EXERCISES

Fill in the blank spaces with the words which most accurately complete the statement.

In the high school low hurdles race, the distance from the start to the first hurdle is \_\_\_\_\_ yards; it is \_\_\_\_\_ yards between the hurdles; and it is \_\_\_\_\_ yards from the last hurdle to the finish. There are \_\_\_\_\_ hurdles to be cleared.

### Tests of proficiency in sports

Numerous articles in the physical education journals present tests of techniques in a variety of sports.<sup>17</sup> These tests are usually based upon an analysis of the skills involved in the sport. Validation of the batteries of tests is by means of comparisons between scores made by pupils and teachers' judgments of pupil proficiency. Ashbrook and his colleagues listed the *Heath-Rodgers Soccer Test* for elementary-school boys, the *Dyer Backboard Test of Tennis Ability*, the *French-Cooper* and *Russell-Lange Volleyball Tests* for high-school girls, and the *Dyer-Schurig-Apgar Basketball Test* for high-school girls as those available for use at the elementary- and junior-high-school levels.<sup>18</sup>

### Physical classification tests

The importance of tools to be used in the classification of pupils for physical education and particularly for competitive sports is obvious. Physical differences among pupils of the same age are so great that classification by chronological age is likely to result in injuries to the smaller and weaker children and usually deprives them of adequate opportunities for exercise. Two indices useful for classification purposes at the elementary- and junior-high-school levels have been validated. As the brief indications of their nature make clear, such indices are obtained by the use of physical rather than paper-and-pencil tests.

<sup>17</sup> See bibliography at end of this chapter for such references.

<sup>18</sup> Ashbrook, Espenschade, and Cozens, *op. cit.* p. 837-38.



McCloy developed a classification index for elementary-school children.<sup>19</sup> The formula is as follows:

$$\text{Classification Index} = 20A + 6H + W,$$

where A refers to age in years, H to height in inches, and W to weight in pounds. Another index making use of the same physical and age measures was derived for junior-high-school girls.<sup>20</sup> The index is obtained by the use of the formula:

$$\text{Index} = 2A + H + .11W.$$

Ashbrook, Espenschade, and Cozens stated that the factors of age, weight, and height when properly combined are probably almost as useful for classification purposes as are more complex measures and the simplicity of their use is of considerable importance.<sup>21</sup>

## 6 DIAGNOSIS IN PHYSICAL EDUCATION

Diagnosis in physical education as well as in health education appears to depend much more upon teacher observation and physical examinations than upon any standardized testing devices of the pencil-and-paper type. The tests of general physical qualities and of physical fitness serve some diagnostic functions. Other tests of diagnostic value are those for the measurement of blood pressure under varying conditions of fatigue. Both of these types can be given by a skilled teacher. Still other tests require technical knowledge and equipment not ordinarily possessed by the teacher.

A significant trend in diagnosis in physical education is that the issue is being approached from the functional rather than the structural standpoint. Even with functional tests, however, it is felt by some that the tests fail to measure the functioning of such organs as the nervous system, for example, completely enough to furnish a highly satisfactory diagnostic score.<sup>22</sup>

<sup>19</sup> C. H. McCloy, *The Measurement of Athletic Power*. A. S. Barnes and Co., New York, 1932.

<sup>20</sup> F. W. Cozens, Hazel J. Cubberley, and N. P. Neilson, *Achievement Scales in Physical Education Activities for Secondary School Girls and College Women*. A. S. Barnes and Co., New York, 1937.

<sup>21</sup> Ashbrook, Espenschade, and Cozens, *op. cit.* p. 837.

<sup>22</sup> Whitelaw R. Morrison and Laurence B. Chenoweth, *Normal and Elementary Physical Diagnosis*. Lea and Febiger, Philadelphia, 1932. p. 331-33.

## Topics for Discussion

1. Discuss the aims of health education.
2. Comment upon the nature and present status of health knowledge testing.
3. What is a major limitation of health attitudes inventories?
4. Discuss some of the preventive and diagnostic procedures in health education for use in the classroom.
5. What are the major objectives of physical education?
6. In what way are some of the measures of general physical qualities useful indications of health status?
7. In what way are cardiovascular and posture tests useful in physical education?
8. Illustrate some methods of testing knowledge and information in physical education.
9. Indicate the nature of tests of proficiency in sports.
10. Give some of the procedures useful in the classification of pupils for physical education.
11. Discuss diagnostic methods in physical education.

## Selected References

- ASHBROOK, WILLARD P., ESPENSCHADE, ANNA, AND COZENS, FREDERICK W. "Physical Education—Measurement." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 835-42.
- BOVARD, JOHN F., AND OTHERS. *Tests and Measurements in Physical Education*. Third edition. Philadelphia: W. B. Saunders Co., 1949.
- BRACE, DAVID K. "The Development of Measures of Pupil Achievement in Physical Education." *Research Quarterly of the American Physical Education Association*, 2:32-37; October 1931.
- BRACE, DAVID K. *Measuring Motor Ability*. New York: A. S. Barnes and Co., 1927.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 543-53, 563-65.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p. 320-24, 333-35.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938. p. 87-88.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*.



- New Brunswick, N. J.: Rutgers University Press, 1949. p. 475-81, 486.
- CARPENTER, AILEEN. "The Measurement of General Motor Capacity and General Motor Ability in the First Three Grades." *Research Quarterly of the American Physical Education Association*, 13:445-65; December 1942.
- CARPENTER, AILEEN. "Strength Testing in the First Three Grades." *Research Quarterly of the American Physical Education Association*, 13:328-32; October 1942.
- CLARKE, H. HARRISON. *The Application of Measurement to Health and Physical Education*. Second edition. New York: Prentice-Hall, Inc., 1950.
- COWELL, CHARLES C. "Evaluation versus Measurement in Physical Education." *Journal of Health and Physical Education*, 12:499-501, 534-35; November 1941.
- CURETON, THOMAS K., AND OTHERS. "The Measurement of Understanding in Physical Education." *The Measurement of Understanding*. Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1946. Chapter 12.
- DYER, JOANNA T. "The Backboard Test of Tennis Ability." *Research Quarterly of the American Physical Education Association*, Supplement, 6:62-74; March 1935.
- DYER, JOANNA T., SCHURIG, JENNIE C., AND APGAR, SARA L. "A Basketball Motor Ability Test for College Women and Secondary School Girls." *Research Quarterly of the American Physical Education Association*, 10:128-47; October 1939.
- FRENCH, ESTHER L. "The Construction of Knowledge Tests in Selected Professional Courses in Physical Education." *Research Quarterly of the American Physical Education Association*, 14:406-24; December 1943.
- FRENCH, ESTHER L., AND COOPER, BERNICE I. "Achievement Tests in Volleyball for High School Girls." *Research Quarterly of the American Physical Education Association*, 8:150-57; May 1937.
- GLASSOW, RUTH B., AND BROER, MARION R. *Measuring Achievement in Physical Education*. Philadelphia: W. B. Saunders Co., 1938.
- GUDAKUNST, DON W. "Diagnosis in Health Education." *Educational Diagnosis*. Thirty-Fourth Yearbook of the National Society for the Study of Education. Bloomington, Ill.: Public School Publishing Co., 1935. Chapter 17.
- HEATH, MARJORIE L., AND RODGERS, ELIZABETH G. "A Study in the Use of Knowledge and Skill Tests in Soccer." *Research Quarterly of the American Physical Education Association*, 3:33-53; December 1932.

- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. Chapter 13.
- MCCLOY, CHARLES H. *Appraising Physical Status: Methods and Norms*. University of Iowa Studies in Child Welfare, Vol. XV, No. 2. Iowa City: University of Iowa, 1938.
- MCCLOY, CHARLES H. *Appraising Physical Status: The Selection of Measurements*. University of Iowa Studies in Child Welfare, Vol. XII, No. 2. Iowa City: University of Iowa, 1936.
- MCCLOY, CHARLES H. *Tests and Measurements in Health and Physical Education*. Second edition. New York: F. S. Crofts and Co., 1942.
- NEILSON, NEILS P., AND COZENS, FREDERICK W. *Achievement Scales in Physical Education Activities: For Boys and Girls in Elementary and Junior High Schools*. New York: A. S. Barnes and Co., 1934.
- ROGERS, FREDERICK R. *Physical Capacity Tests*. New York: A. S. Barnes and Co., 1938.
- RUGEN, MABEL E., AND NYSWANDER, DOROTHY. "The Measurement of Understanding in Health Education." *The Measurement of Understanding*. Forty-Fifth Yearbook of the National Society for the Study of Education, Part I. Chicago: University of Chicago Press, 1946. Chapter 11.
- RUSSELL, NAOMI, AND LANGE, ELIZABETH. "Achievement Tests in Volleyball for High School Girls." *Research Quarterly of the American Physical Education Association*, 11:33-41; December 1940.
- SCHROEDER, E. M. *On Measurement of Motor Skills: An Approach through a Statistical Analysis of Archery Scores*. New York: King's Crown Press, 1945.
- SCOTT, M. GLADYS, AND FRENCH, ESTHER L. *Better Teaching through Testing*. New York: A. S. Barnes and Co., 1945.
- STRANG, RUTH. "Health Education." *Encyclopedia of Educational Research*. Revised edition. New York: Macmillan Co., 1950. p. 529-40.



## ***Measuring and Evaluating General Educational Achievement***

THIS CHAPTER treats the following points in the measurement and evaluation of general educational achievement:

- A. Advantages of general achievement batteries.
- B. Limitations of measures of general achievement.
- C. General *vs.* specific surveys.
- D. Types of achievement batteries.
- E. Some distinctive features of certain achievement batteries.

The emphasis throughout this volume is rather definitely on diagnostic and analytic testing and on evaluative techniques in subject and performance areas. However, a consideration of the practical problems of measurement and evaluation in the classroom leads to the conviction that there is a real service to be rendered by survey tests of general achievement. Accordingly, tests of that type are treated briefly here.

### **1 MEASUREMENT OF GENERAL ACHIEVEMENT**

#### **General *vs.* specific measurement**

The battery type of general achievement test opens up certain types of possibilities for diagnostic, analytic, and remedial work and for the use of test results in educational guidance. Such a test affords a rather complete survey of the pupil's educational status. It pre-

sents a perspective of the aspects of his accomplishment measurable by paper-and-pencil tests.

A general survey test of the types described in this chapter may reveal a specific weakness in, for example, the language skills. To the critical teacher this is a challenge to discover more exactly the factors underlying the deficiency. Accordingly, the cases identified by the test as weak in certain areas should be subjected to a detailed analytic or diagnostic test in the subject for the purpose of locating specific difficulties and their causes.

Results from general achievement tests are of further value in affording one basis for educational guidance of pupils. Evidence concerning the suitability for the individual pupil of certain subsequent courses or programs and even of certain vocations may often be obtained early in the pupil's school career. When cumulated over a period of years and when supplemented by other evidence, such test results can contribute significantly to pupil guidance.

### Types of general achievement batteries

There are available today a number of general achievement batteries, several of which have very distinct merit. Most of the better-known and more widely used batteries are briefly described here. No attempt is made to illustrate their measurement techniques, for the wide variety of outcomes tested makes that impracticable. Moreover, illustrations from some of these tests appear in preceding chapters of this volume. The batteries discussed and summarized below are classified under four headings: (1) general achievement batteries, (2) achievement batteries in skill areas, (3) achievement batteries in content areas, and (4) specialized achievement batteries.

### Advantages and limitations of general achievement tests

Among the specific qualities of the battery type tests of general achievement that have been given considerable emphasis by persons interested in the improvement of classroom measurement are the following:

*Comparable units of measurement.* The use of a uniform unit of measurement in the scaling of battery tests constitutes a real advantage in the interpretation of the test results and in the comparisons of results from one subject field to another. While this is an im-



portant advantage, it does not at all mean that uniformity in units of measurement may not be secured in single tests in unrelated subjects.

*Unity of population in standardization.* The fact that the standardization of most comprehensive batteries is based upon results from the same pupils for each of the different subject tests insures a better picture of the relationships of achievement in these different subjects. For example, the reading achievement of pupils of a certain grade can be compared with their language achievement only when tests are standardized under these conditions.

*Simplicity of interpretation.* The use of comparable units of measurement and similar testing techniques in the several tests comprising a general achievement battery simplifies the problems of comparing and interpreting the results. The raw test scores are readily turned into standard scores, educational ages, and grade equivalents. Modern graphic methods of summarizing test results make effective use of such derived scores. Profile charts of the type commonly provided with these tests add to the clearness with which test results may be interpreted. Naturally, such profiles are useful only in case test scores from a number of different tests are reducible to a common unit of measurement.

*Ease of administration and scoring.* The tendency of the authors of battery type tests to utilize the same or similar types of testing techniques throughout the battery unquestionably tends to simplify the problems of administering it. The use of uniform methods of recording the pupil's responses also simplifies the problem of scoring. In general, however, such battery tests are usually so long that the time required to administer them and the labor involved in scoring them become quite great. However, it may be that this is not too high a price to pay for extensive sampling and valid and reliable measurement. Furthermore, many of these tests are now available for either hand- or machine-scoring above the primary grades.

*Economy of cost.* Any economy that results from the use of battery tests appears to be conditioned by the assumption that broadly diagnostic rather than specifically diagnostic or analytic measurement is desired. It is probably true that almost any one of the modern batteries of achievement tests will furnish a wider sampling into more subject fields at a lower cost per pupil than could be accomplished by the selection of single-subject tests for the purpose. There are numerous occasions, however, when it is of greater impor-

tance to measure more intensively a limited range of subjects. For this type of measurement the battery tests are usually not the most economical. In order to provide for this situation, the authors of most test batteries have prepared the tests for certain subjects in separate form. Indeed, several of the battery tests are available only in the form of separate coordinated test booklets.

## 2 GENERAL ACHIEVEMENT BATTERIES

More than half of the achievement test batteries now available are designed for measuring outcomes in all of the major instructional areas appropriate for the grade levels in question. As instructional emphases in the primary grades are mainly upon skill in reading, arithmetic, and language, battery tests designed for use in the first three grades concentrate upon these skill areas almost entirely. Such content subjects as the social studies and elementary sciences commonly receive more attention in the intermediate and upper grades. For this reason, batteries designed for use above Grade 3 usually include parts in these subject areas and often on health and safety as well as sections in the areas of receptive and expressive language and computational skills. Consequently, the number of separate parts and of resulting scores tends to be greater at the intermediate and upper grade levels than for the primary grades. There is a parallel tendency for time requirements in administration and scoring to be greater at the intermediate and upper grade levels than for the primary grades.

### Stanford Achievement Tests

The original battery of these tests, published in 1923, was one of the outstanding measuring instruments of that period. The tests set new standards of validity, reliability, and other examination criteria for later workers and undoubtedly did much to stimulate the improvement of educational measurement in general. After six years, and on the basis of much critical analysis and experimentation, the tests were revised in the form known as the *New Stanford Achievement Tests*. They have more recently been revised a second and a third time and are again known as the *Stanford Achievement Tests*.

In their present form the batteries consist of five primary tests for Grades 2 and 3, six elementary tests for Grades 3 and 4, nine



intermediate tests for Grades 5 to 7, and nine advanced tests for Grades 7 to 9. The testing times are 85 minutes for the Primary battery, 145 minutes for the Elementary battery, and 232 and 227 minutes respectively for the Intermediate and Advanced batteries. At the two upper levels the tests are also available for machine scoring in partial batteries. Norms are of two types: (1) percentile norms by grades, and (2) modal-age norms, based only on those pupils in the standardization group who were at grade for their age.

TABLE 37. Summary of tests and timing: Stanford Achievement Tests <sup>1</sup>

Area	Test	Level, Grades, and Timing			
		Pri- mary	Ele- mentary	Inter- mediate	Ad- vanced
		2-3	3-4	5-7	7-9
Language Arts	Reading: Paragraph Meaning	25	30	30	30
	Word Meaning	10	10	12	12
	Spelling	15	20	15	15
Arithmetic	Language		25	20	20
	Arithmetic Computation	18	30	40	40
	Reasoning	17	30	35	35
Social Studies	Social Studies			20	20
Science	Science			15	15
Study Skills	Study Skills			45	40

## Metropolitan Achievement Tests

The present batteries of these tests represent a second revision. The original edition was published in the 1920s and the first revision was issued in several forms during the period 1931 to 1937. Four tests in reading and arithmetic appear in the Primary I battery for Grades 1 and 2, whereas the Primary II battery for Grades 2 and 3 includes five tests and extends to the language arts as a third area. The Elementary battery for Grades 3 to 5 consists of six tests in these same three basic skills areas. Ten tests appear in the Intermediate battery for Grades 5 to 7 and in the Advanced battery for Grades 7 to 9, including tests in the two content areas of social studies and sciences. The working time required for the various batteries is 45 minutes for the Primary I, 85 minutes for the Primary II, 135 minutes for the Elementary, 215 minutes for the Inter-

<sup>1</sup> Truman L. Kelley and others, *Stanford Achievement Tests*, Primary, Elementary, Intermediate, and Advanced. World Book Co., Yonkers, N. Y., 1953.

mediate, and 225 minutes for the Advanced. Each battery is issued in a single-booklet edition in Forms R, S, T, and U. The reading tests and the arithmetic tests are also available in separate booklets for the three batteries above the primary level.

TABLE 38. Summary of tests and timing: Metropolitan Achievement Tests <sup>2</sup>

Area	Test	Level, Grades, and Timing				
		Primary		Elementary	Intermediate	Advanced
		I	II			
		1-2	2-3	3-5	5-7	7-9
Reading	Word Picture	15				
	Word Recognition	10				
	Word Meaning	10	10			
	Reading		30	25	25	25
	Vocabulary			10	10	10
Arithmetic	Literature				15	15
	Numbers	10				
	Arithmetic Fundamentals		15	35	40	40
	Arithmetic Problems		15	30	40	40
Language	Spelling		15	15	15	15
	Language Usage			20		
	English				25	35
Social Studies	History				15	15
	Geography				15	15
Science	Science				15	15

Grade and age norms of both the traditional and the modal-age types are furnished for each test at each level. Percentile grade norms of both the traditional and modal-age types are also provided for each test at each level for October 16-November 15, January 16-February 15, and April 16-May 15 testing dates. Traditional grade norms are also furnished for pupils in parochial schools and for Negro pupils in segregated schools.

### Coordinated Scales of Attainment

These scales, successors to the *Unit Scales of Attainment*, are issued in eight separate single-booklet editions for Grades 1 to 8 in Forms A and B. Batteries for the primary grades entail the recording of answers in the booklets and must be scored manually.

<sup>2</sup> Gertrude H. Hildreth, Richard D. Allen, and others, *Metropolitan Achievement Tests*, Primary, Elementary, Intermediate, and Advanced. World Book Co., Yonkers, N. Y., 1946.



Pupils in Grades 4 to 8 record their answers on an answer sheet which may be scored either by hand or by machine. Working times for pupils vary from approximately 100 minutes in the primary grades to 256 minutes for the intermediate and upper grades. Tests are available in separate booklets for the various subject areas. Norms are provided in the form of grade equivalents and age equivalents for major scores.

TABLE 39. Summary of tests and timing: Coordinated Scales of Attainment<sup>3</sup>

Area	Test	Battery and Grade							
		1	2	3	4	5	6	7	8
Reading	Picture-Word Association	10	10	10					
	Word-Picture Association	10	10	10					
	Vocabulary Recognition	10	10	10					
	Reading Comprehension	10	10	10					
Arithmetic	Reading				45	45	45	45	4
	Reading Experience—Literature				15	15	15	15	15
	Arithmetic Experience	10	10						
	Number Skills	10	10						
	Arithmetic Computation	15	15	20	45	45	45	45	45
Language	Arithmetic Problem Reasoning	15	15	20	30	30	30	30	30
	Spelling		20	20	40	40	40	40	40
	Punctuation				12	12	12	12	12
	Capitalization				12	12	12	12	12
Social Studies	Usage				12	12	12	12	12
	History				15	15	15	15	15
	Geography				15	15	15	15	15
Science	Elementary Science				15	15	15	15	15

## Modern School Achievement Tests

This single-booklet battery, issued in Forms I and II, includes ten tests in six subject areas. The content ranges in difficulty from material suitable for use in Grade 2 to that appropriate for pupils in Grade 8. The total testing time is slightly over 160 minutes. Age and<sup>3</sup> grade norms are provided for each of the ten tests and a table is provided for the interpretation of percentage of accuracy scores on the reading speed and accuracy test in terms of grade equivalents.

<sup>3</sup> M. E. Branom and others, *Coordinated Scales of Attainment*, Grades 1, 2, 3, 4, 5, 6, 7, and 8. Educational Test Bureau, Minneapolis, 1946.

TABLE 40. Summary of parts and timing: Modern School Achievement Tests <sup>4</sup>

Area	Part	Grades and Timing	
		2-5	6-8
Reading	Reading: Level of Comprehension	30	30
	Reading: Speed and Accuracy	8	5
Arithmetic	Arithmetic Computation	20	20
	Arithmetic Reasoning	25	25
Language	Language Usage	10	10
	Spelling	15	15
Social Studies	History and Civics	15	15
	Geography	15	15
Science	Elementary Science	12	12
Health	Health Knowledge	15	15

### Gray-Votaw-Rogers General Achievement Tests

A revision of the *Gray-Votaw General Achievement Tests*, these 1951 editions appear at the Primary, Intermediate, and Advanced

TABLE 41. Summary of tests and timing: Gray-Votaw-Rogers General Achievement Tests <sup>5</sup>

Area	Test	Level, Grades, and Timing			
		Primary		Inter- mediate	Ad- vanced
		1-2	3	4-6	7-9
Reading	Comprehension	15	12	10	10
	Vocabulary	10	5	8	8
	Literature			12	12
Arithmetic	Computation	10	10	28	28
	Reasoning	12	8	20	20
Language	Language			10	10
	Spelling	15	15	15	15
Social Studies	Social Studies			12	12
Science	Elementary Science			10	10
Health and Safety	Health and Safety			10	10

<sup>4</sup> Arthur I. Gates and others, *The Modern School Achievement Tests*. Bureau of Publications, Teachers College, Columbia University, New York, 1931.

<sup>5</sup> Hob Gray, David F. Votaw, and J. Lloyd Rogers, *Gray-Votaw-Rogers General Achievement Tests*, Primary, Intermediate, and Advanced. Steck Co., Austin, Texas, 1950-51.



levels and provide tests from Grade 1 to Grade 9. Forms Q, R, S, and T are available at each level. Pupil working time on the Primary battery is 62 minutes in Grades 1 and 2 and 50 minutes in Grade 3. Working time is 135 minutes each on the higher-level batteries. These batteries are hand-scorable, but an Abbreviated Edition, available in Forms U, V, W, and X for Grades 5 to 9, is accompanied by machine-scorable answer sheets that can also be scored manually. Grade and age norms are provided for each test and for a total score on each battery and percentile grade norms are given for the total battery score at each level.

### Master Achievement Tests

The six levels of this battery are for use in Grades 3 to 8. Two forms, A and B, are available in single booklets for each grade. The testing time in Grades 7 and 8 is about 140 minutes. In the four lower grades the time requirement is slightly longer. Grade norms, by means of which grade equivalents, or G-scores, may be obtained, are provided for the various tests of the battery at each grade level.

TABLE 42. Summary of tests and timing: Master Achievement Tests <sup>6</sup>

Area	Test	Grades and Timing					
		3	4	5	6	7	8
Reading	Reading	30	23	18	16	12	11
	Vocabulary	15					
Arithmetic	Arithmetic	60	57	56	56	50	64
Language	English	30	26	18	23	26	32
	Spelling	?	?	?	?		
Social Studies	Geography		20	20	23	23	
	History			16	15	15	15
Science and Health	Science and Health	24	15	15	12	15	17

### National Achievement Tests

These tests are issued in two batteries, for the intermediate grades and for the upper grades and junior high school. The testing time is slightly over 200 minutes for each level. Three sets of grade

<sup>6</sup> *Master Achievement Tests*, Grades 3, 4, 5, 6, 7, and 8. American Education Press, Columbus, Ohio, 1937.

equivalents and age equivalents are furnished for each of the ten tests—for pupils having *IQs* below 90, between 90 and 109, and above 109.

TABLE 43. Summary of tests and timing: National Achievement Tests <sup>7</sup>

Area	Test	Part	Grades and Timing	
			3-6	6-8
Reading	Reading Comprehension	Following Directions Sentence Meaning Paragraph Meaning	30	30
Spelling Arithmetic	Reading Speed		3	2
	Spelling		15	15
	Arithmetic Fundamentals	Computation Number Comparisons	30	30
	Arithmetic Reasoning	Comparisons Problem Analysis Problems	30	30
Language	English	Language Usage—Words Language Usage—Sentences Punctuation and Capitalization	30	30
Literature	Literature	Expressing Ideas Motives and Moods Miscellaneous Facts	20	20
Social Studies	Geography	Geographical Ideas and Comparisons Miscellaneous Facts	20	20
	History and Civics	Lessons of History Historical Facts	15	15
Health	Health		10	10

## Cooperative Achievement Tests

This series of junior-high-school tests is published in four and in six booklets, as the English and reading tests are issued both in a one-booklet edition and as three separate booklets. Separate answer sheets for machine- or hand-scoring are available. Several forms of each test are now available and new forms are issued periodically. End-of-year percentile grade norms based on scaled scores are furnished for pupils in several different types of school organizations.

<sup>7</sup> Robert K. Speer and Samuel Smith, *National Achievement Tests*, Municipal Battery. Acorn Publishing Co., Rockville Centre, N. Y., 1938.



TABLE 44. Summary of tests and timing: Cooperative Achievement Tests for the junior high school <sup>8</sup>

Area	Test	Part	Timing
English	Mechanics of Expression, A	Grammatical Usage	15
		Punctuation and Capitalization	15
	Effectiveness of Expression B <sub>1</sub>	Spelling	10
		Sentence Structure and Style	15
		Diction	10
Reading	Reading Comprehension, C <sub>1</sub>	Organization	15
		Vocabulary	15
		Speed of Comprehension	25
		Level of Comprehension	
Mathematics	Mathematics for Grades 7, 8, and 9	Skills	30
		Facts, Terms, and Concepts	10
		Applications	30
		Appreciation	10
Science	Science for Grades 7, 8, and 9	Informational Background	40
		Terms and Concepts	15
		Comprehension and Interpretation	25
Social Studies	Social Studies for Grades 7, 8, and 9	Informational Background	40
		Terms and Concepts	15
		Comprehension and Interpretation	25

### 3 ACHIEVEMENT BATTERIES IN SKILL AREAS

A second group of test batteries is distinguishable from those discussed above by the fact that they are concerned with the skill areas to the exclusion of the content areas of instruction. These batteries are not restricted to the primary grade level, where the instructional emphasis is mainly upon skills, but variously provide tests for the intermediate and upper grades and for the entire range from primary to junior-high-school grades.

#### Iowa Every-Pupil Tests of Basic Skills

Growing out of the basic skills tests used in a state-wide testing program in Iowa over a period of years, the present editions of this battery cut across traditional subject lines in at least one test. Tests A, C, and D are designed for the measurement of reading, language, and arithmetic skills respectively, but Test B, Work-Study Skills, has no direct counterpart in the typical program of studies. The tests

<sup>8</sup> *Cooperative Achievement Tests*. Cooperative Test Division, Educational Testing Service, Princeton, N. J., 1947-50.

are issued in four forms—L, M, N, and O—and as an Elementary battery for Grades 3 to 5 and an Advanced battery for Grades 5 to 9. Separate answer sheets for machine-scoring are available for the Advanced battery. The pupil working time is 196 minutes at the lower level and 268 minutes at the upper level.

TABLE 45. Summary of tests and timing: Iowa Every-Pupil Tests of Basic Skills<sup>9</sup>

Test	Part	Level, Grades, and Timing	
		Elementary	Advanced
		3-5	5-9
Silent Reading Comprehension Work-Study Skills	Reading Comprehension	36	58
	Vocabulary	10	10
	Map Reading	11	28
	Use of References	8	5
	Use of Index	8	10
	Use of Dictionary	12	17
	Alphabetization	8	
	Reading Graphs, Charts, and Tables		17
Basic Language Skills	Punctuation	11	17
	Capitalization	8	10
	Usage	13	16
	Spelling	8	12
	Sentence Sense	6	
Basic Arithmetic Skills	Vocabulary and Fundamental Knowledge	12	15
	Fundamental Operations	20	25
	Problems	25	28

Both pupil norms and school norms are provided with these test batteries. Pupil norms for each part and each test are of three types: (1) grade norms, (2) age-at-grade norms, and (3) percentile grade norms. School norms of the type discussed in Chapter 5 are also furnished for each test part.

## California Achievement Tests

The *Progressive Achievement Tests* have been retitled *California Achievement Tests* in the new 1951 editions. Coverage of Grades 1 to 9 is provided by the primary, elementary, and intermediate levels of the battery. The six tests are published in three separate booklets

<sup>9</sup> E. F. Lindquist, editor, *Iowa Every-Pupil Tests of Basic Skills*, Elementary and Advanced. Houghton Mifflin Co., Boston, 1940-43.



at the two higher levels, with one booklet each for reading, arithmetic, and language, and in single-booklet editions at all three levels. Four forms—AA, BB, CC, and DD—are available at each level. Working times for pupils are 95, 119, and 151 minutes at the primary, elementary, and intermediate levels, respectively, when answers are recorded in the test booklets and slightly more when, as is optional for levels above the primary, machine-scored answer sheets or the *CTB Scoreze* answer sheets are used.

TABLE 46. Summary of tests and timing: California Achievement Tests<sup>10</sup>

Area	Test	Part	Level, Grades, and Timing		
			Pri- mary	Ele- mentary	Inter- mediate
			1-4	4-6	7-9
Reading	Vocabulary	Word Form	5	3	
		Word Recognition	4	3	
		Meaning of Opposites	5	3	
		Meaning of Similarities		3	
		Mathematics			3
	Comprehension	Science			3
		Social Science			3
		General			3
		Following Directions	5	5	8
		Directly Stated Facts	5		
Arithmetic	Reasoning	Interpretations	5	12	25
		Reference Skills		6	5
		Number and Sequence	2		
		Money	2		
		Number and Time	4		
	Fundamentals	Signs and Symbols	4	3	
		Problems	10	10	16
		Number Concept		3	4
		Symbols and Rules			5
		Numbers and Equations			5
		Addition	5	10	10
		Subtraction	7	10	10
		Multiplication	8	12	12
		Problems	8		
		Division		12	12
Language	Mechanics of English	Capitalization	3	3	3
		Punctuation	3	4	4
		Words and Sentences		5	4
	Spelling	Parts of Speech			6
		Spelling	10	12	10

<sup>10</sup> Ernest W. Tiegs and Willis W. Clark, *California Achievement Tests*, Primary, Elementary, and Intermediate. California Test Bureau, Los Angeles, 1950.

Grade and age equivalents and percentile grade norms are provided for each of the six tests, for total achievement separately in reading, arithmetic, and language, and also for a total score on the complete battery. A handwriting test is provided for use if desired and a scale for interpreting quality in terms of grade placement appears in the manuals for the language test. However, age equivalents and percentile norms are not provided for handwriting, and this test is not an integral part of the battery.

### American School Achievement Tests

These tests are available in single-booklet form for each of four levels ranging from Grade 1 to Grade 9. The Primary I battery appears in Forms A and B but Forms A, B, and C are available for the three other batteries. Pupil working time varies from 50 and 65 minutes for the Primary I and II batteries to 107 minutes for the Intermediate and 127 minutes for the Advanced batteries. Machine-scoring answer sheets are available for the Intermediate and Advanced batteries. Norms are provided in the form of age equivalents and grade equivalents at all four levels.

TABLE 47. Summary of tests and timing: American School Achievement Tests<sup>11</sup>

Area	Test	Level, Grades, and Timing			
		Primary		Inter- mediate	Ad- vanced
		I	II		
		1	2-3	4-6	7-9
Reading	Word Recognition	5			
	Word Meaning	10			
Arithmetic	Sentence and Word Meaning		10	10	10
	Paragraph Meaning		15	15	20
	Numbers	35			
	Computation		12	30	35
Language	Problems		10	20	25
	Language		8	20	25
	Spelling		10	12	22

<sup>11</sup> Robert V. Young and others, *American School Achievement Tests*, Primary I and II, Intermediate, and Advanced. Public School Publishing Co., Bloomington, Ill., 1941-47.



## Modern School Achievement Tests

These tests are issued not only in the complete edition discussed and summarized in the preceding section of this chapter but also in a single-booklet Short Form covering the skill subjects of reading, arithmetic, and language. Since the characteristics of the complete battery apply to the short form, no further discussion of this edition is necessary here.

### 4 ACHIEVEMENT BATTERIES IN CONTENT AREAS

General achievement batteries dealing with the content aspects of subject areas are considerably less numerous than those in the basic skills fields. In fact, the authors know of only one instrument of this type that may properly be considered a general achievement battery. Tests that deal broadly with subject fields or areas are considered in Chapters 15 to 21, but there remains for brief mention here the battery that deals with the content of three related content areas.

TABLE 48. Summary of tests and timing: Progressive Tests in Social and Related Sciences <sup>12</sup>

Area	Test	Section	Timing
Social Studies I	The American Heritage	Exploration and Colonization The Westward Movement Later Development of the Nation Understanding of Democracy	50
	Peoples of Other Lands and Times	Peoples of Other Lands	
Social Studies II	Geography	Peoples of Other Times Geographic Facts: The United States Geographic Facts: The World Reading Maps: Knowledge of Geographic Terms	60
	Basic Social Processes	Effects of Geography on the Life of Man Food, Clothing, and Shelter Transportation and Communication	
Related Sciences III	Health and Safety	Eating for Health Other Health Information Safety Information	40
	Elementary Science	The World about Us Man's Increasing Control over Nature	

<sup>12</sup> Georgia S. Adams and John A. Sexson, *Progressive Tests in Social and Related Sciences*, Elementary. California Test Bureau, Los Angeles, 1947.

## Progressive Tests in Social and Related Sciences

This series of six tests in three booklets includes four tests in the social studies and two tests in related sciences. The battery, available in Forms A and B, requires a minimum of 150 minutes of testing time when answers are recorded in the booklets and a time allowance greater by approximately one-fourth when separate answer sheets are used. The separate answer sheets can be scored either by machine or by hand. Age and grade norms are furnished for each of the six tests and percentile norms in skeleton form are provided for pupils in Grades 4 to 8 for each test and each section.

## 5 SPECIALIZED BATTERIES

Two specialized batteries involving the measurement of general achievement are the *Comprehensive Test Program*<sup>13</sup> and the *Otis Classification Test*.<sup>14</sup> The former, for use in Grades 4 to 9, includes an intelligence test, an educational background questionnaire, a school practices questionnaire, and a comprehensive achievement test. The latter consists of two parts measuring general intelligence and general achievement.

## Topics for Discussion

1. What major functions are served by batteries of general achievement tests?
2. What are some of the advantages of general achievement test batteries?
3. What are some of the limitations of general achievement test batteries?
4. How do the functions of these test batteries and of analytic or diagnostic tests differ?
5. What are some of the instructional outcomes achievement test batteries fail to measure?
6. How may general achievement tests be used in the grade placement and sectioning of pupils in a school system?
7. Into what major types may achievement test batteries be classified?

<sup>13</sup> William A. McCall and John P. Herring, *A Comprehensive Test Program: Manual for Teachers*. Laidlaw Brothers, Inc., Chicago, 1937.

<sup>14</sup> Arthur S. Otis, *Otis Classification Test*. World Book Co., Yonkers, N. Y., 1941.



8. What are the advantages of the pupil profile charts provided with most achievement test batteries?
9. How can achievement test batteries and pupil profile charts be used in the measurement of the educational progress of pupils?

### Selected References

- ADAMS, GEORGIA S., AND SEXSON, JOHN A. *Manual of Directions: Progressive Tests in Social and Related Sciences*, Elementary Battery. Los Angeles: California Test Bureau, 1947.
- BUROS, OSCAR K., editor. *The Fourth Mental Measurements Yearbook*. Highland Park, N. J.: Gryphon Press, 1953. p. 1-66.
- BUROS, OSCAR K., editor. *The Nineteen Forty Mental Measurements Yearbook*. Highland Park, N. J.: Mental Measurements Yearbook, 1941. p. 19-49.
- BUROS, OSCAR K., editor. *The Nineteen Thirty Eight Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1938. p. 14-33.
- BUROS, OSCAR K., editor. *The Third Mental Measurements Yearbook*. New Brunswick, N. J.: Rutgers University Press, 1949. p. 1-50.
- The Cooperative Achievement Tests: A Handbook Describing Their Purpose, Content, and Interpretation*. New York: Cooperative Test Service, October 1936.
- Coordinated Scales of Attainment: (1) Master Manual, Batteries 1, 2, and 3, (2) Master Manual, Batteries 4 through 8, and (3) Guide to Remedial Work, Batteries 1-8*. Minneapolis: Educational Test Bureau, 1950.
- GATES, ARTHUR I., AND OTHERS. *Modern School Achievement Tests: Manual of Directions*. New York: Bureau of Publications, Teachers College, Columbia University, 1931.
- HILDRETH, GERTRUDE H. *Metropolitan Achievement Tests: Manual for Interpreting*. Yonkers, N. Y.: World Book Co., 1948.
- JORDAN, A. M. *Measurement in Education*. New York: McGraw-Hill Book Co., Inc., 1953. Chapter 4.
- Manual for Interpretation of Iowa Every-Pupil Tests of Basic Skills*, Form O. Boston: Houghton Mifflin Co., 1943.
- Manual of Directions and Interpretations: Gray-Votaw-Rogers General Achievement Tests*, (1) Primary, and (2) Intermediate and Advanced. Austin, Texas: Steck Co., 1948.
- MCCALL, WILLIAM A. *Measurement*. New York: Macmillan Co., 1939. Chapters 14-15.

- McCALL, WILLIAM A., AND HERRING, JOHN P. *A Comprehensive Test Program: Manual for Teachers*. Chicago: Laidlaw Brothers, Inc., 1937.
- NELSON, M. J. *Tests and Measurements in Elementary Education*. New York: Cordon Co., 1939. Chapter 10.
- TIEGS, ERNEST W., AND CLARK, WILLIS W. *Manual for California Achievement Tests: Complete Battery*, (1) Primary, (2) Elementary, and (3) Intermediate. Los Angeles: California Test Bureau, 1951.
- WEBB, L. W., AND SHOTWELL, ANNA M. *Testing in the Elementary School*. Revised edition. New York: Farrar and Rinehart, Inc., 1939. Chapter 19.
- YOUNG, ROBERT V., AND OTHERS. *Teacher's Manual for American School Achievement Tests*, (1) Primary I, (2) Primary II, (3) Intermediate, and (4) Advanced. Bloomington, Ill.: Public School Publishing Co., 1941-47.



## Glossary

- ability.** The capacity or power to produce.
- accomplishment.** See *achievement*.
- accomplishment quotient.** See *achievement quotient*.
- accuracy.** The ratio between the number of items answered correctly and the number of items attempted.
- achievement.** The accomplishment or production of the pupil in his school work.
- achievement age.** A pupil's level of accomplishment in a particular school subject or field.
- achievement quotient (AQ).** The ratio between educational age and mental age.
- achievement test.** A test that measures the pupil accomplishment resulting from instruction and learning.
- adequacy.** An examination criterion indicating the degree to which a test samples extensively or widely over the content or activities to be tested.
- adjustment.** The process of effecting a satisfactory adaptation to one's environment.
- adjustment inventory.** An instrument used to determine how satisfactorily the individual has adapted himself to his environment.
- administrability.** An examination criterion indicating the characteristics of a test that make for ease and accuracy in giving it.
- age-at-grade norms.** Norms based on pupils grouped or classified by ages within their school grades.
- age equivalent.** The score derived from age norms on a standardized test.
- age norms.** Tables of values representing typical or average performance on standardized tests for pupils in different age groups.
- alternate-response item.** A type of test item to which the pupil responds by indicating which of the two possible answers is right and which is wrong.
- ambiguity.** The quality of a test item that makes possible more than one logical interpretation of its intent or meaning.
- analogies test.** A test of logical reasoning ability involving similarities and dissimilarities.

- analysis.** The process of reducing or taking apart a total performance in the identification of specific skills.
- analytic test.** A test that furnishes a basis for the analysis of skills underlying a performance by securing different measures of abilities contributing to total performance.
- anecdotal record.** An objective account of pupil behavior made by the teacher or some other person observing a significant event in the life of the pupil.
- answer sheet.** A separate piece of paper, usually printed, on which the pupil records his responses for a test.
- applications.** An instructional or learning outcome involving the use of skills, knowledges, concepts, and understandings in practical situations.
- appraise.** See *evaluate*.
- appreciation.** An instructional or learning outcome involving a judgment concerning the worth of a piece of art, an event, or an experience.
- aptitude.** An ability in a certain field or area of performance.
- aptitude test.** A test of specific intelligence, i.e., intelligence as it operates in a certain field or area or performance.
- arithmetic mean (A.M.).** The point on the scale above which and below which the sums of the deviations are equal; the sum of the scores divided by their number.
- array.** A collection of data arranged in a systematic order.
- association method.** A technique of personality evaluation involving free responses to certain stimuli.
- assumed mean.** The midpoint of the class interval in which it is "guessed" that the arithmetic mean will fall.
- attitude.** An instructional or learning outcome represented by a state of readiness which exerts a directive, and sometimes a compulsive, influence upon an individual's behavior.
- attitudes scale.** An instrument used in the determination of pupil opinions or beliefs on an issue or issues which may be controversial in nature.
- average.** A generic term for measures of central tendency.
- basic skills.** Tool skills, such as those of reading, language, and arithmetic.
- basic skills test.** An achievement test measuring performance in such types of communication as speaking, listening, reading, writing, and computing.
- behavior.** All types of responses made by the individual, particularly those that can be observed.



**best answer item.** A type of multiple-choice item to which the pupil responds by attempting to select the best answer from alternatives of which more than one may be correct.

**bi-factor test.** A type of intelligence test from the use of which two scores for separate aspects of mental ability are obtained.

**capacity.** The power to learn or profit from experience.

**case study.** A comprehensive approach to the evaluation of the total personality of the individual pupil.

**central tendency.** A term corresponding to average, commonly applied to the arithmetic mean, median, and midmeasure.

**"chance-half" coefficient.** An estimate of test reliability useful when only one form of a test is available.

**check list.** A list of steps in performing a certain operation used by an observer in evaluating pupil proficiency in some skill.

**chronological age (CA).** Life age; the number of years since birth.

**class analysis chart.** A device for the graphical representation of class performance and individual pupil performance on the various parts of certain achievement tests.

**classification.** The process of assigning a pupil to the grade or unit of a school for which his abilities and training best fit him.

**class interval (c.i.).** One of the divisions of a frequency distribution.

**classroom test.** A test made by the teacher or within a school system for use in specific classes.

**clues.** Characteristics of test items which frequently aid the pupil in determining the correct answers.

**coefficient of alienation (k).** An index of the degree to which two variables are unrelated.

**coefficient of correlation (r).** A measure of relationship that ranges in value from  $+1.00$  through zero to  $-1.00$ ; refers here mainly to Pearson product-moment coefficient.

**comparability.** An examination criterion indicating the characteristic of a test that enables the user to obtain from different administrations of the test results that have equivalent meaning.

**comparable measures.** Scores or values that are expressed in terms of the same unit and with respect to the same point of origin.

**completion exercise.** A type of test exercise to which the pupil responds by filling the blanks of a statement with the words, numbers, or phrases he believes will correctly complete the meaning. •

**composite score.** A single value used to express the results obtained from the use of several different measures.

**comprehension score.** A score indicating the degree of a pupil's understanding of an exercise or of material read.

- concepts.** An instructional or learning outcome involving comprehension of meaning.
- constant error.** A type of deviation from complete accuracy that results from the tendency of some scorers to give high marks and of other scorers to give low marks consistently.
- content subjects.** Fields in which mastery consists mainly in the acquisition of informations and attitudes, as the social sciences and sciences.
- correction.** An adjustment used in computing the arithmetic mean, standard deviation, and correlation coefficient by the short method.
- correction for chance.** A practice followed in scoring some types of objective tests to take account of guessing.
- corrective teaching.** Steps taken to remedy observed defects or difficulties in pupil learning.
- correlation.** The degree of relationship existing between two or more sets of measures.
- correlation chart.** A two-way or double-entry table that shows the relationship existing between pairs of measures for the same individuals or items.
- correlation coefficient.** See *coefficient of correlation*.
- criterion.** A standard by which a test or other product is judged or evaluated.
- cumulative frequency.** The sum of all the scores in a frequency distribution up to any given point.
- cumulative frequency distribution.** A distribution of cumulative frequencies.
- cumulative frequency graph.** A graphical representation of a cumulative frequency distribution.
- cumulative pupil record.** A comprehensive, cumulative record of pupil background, ability, achievement, and behavior.
- curricular validity.** Evidence of test validity shown by adequate coverage of curriculum content by a test.
- cursive writing.** Handwriting with the letters joined.
- decile.** One of the nine points that divide a distribution into ten equal areas.
- derived score.** A value having comparable meaning for the results from various tests.
- deviation.** The amount by which a score or other measure differs from the central tendency of the group of scores in which it is included.
- diagnosis.** The identification and location of specific strengths or weaknesses in performance.



- diagnostic test.** A test used to locate the nature, and if possible the causes, of disability in performance.
- differential aptitude tests.** A term often applied to multiple-factor tests of mental ability.
- difficulty.** The characteristic in a test item that results in a small percentage of correct responses.
- directed observation.** A technique of personality study involving observation of certain specific types of behavior in the pupil.
- discriminative power.** The quality of a test item that results in adequate distinctions in percentages of correct answers by pupils of varying ability levels.
- dispersion.** See *variability*.
- double-entry table.** See *correlation chart*.
- drill test.** A paper-and-pencil instrument designed for use by pupils in practicing certain skills.
- duplicate forms.** See *equivalent forms*.
- economy.** An examination criterion indicating the cost of a test in time and money requirements.
- educational age (EA).** A pupil's level of accomplishment in a number of school subjects.
- educational quotient (EQ).** The ratio between educational age and chronological age.
- educational test.** A measuring instrument that appraises the results or effects of instruction and learning.
- elementary mathematics.** Primarily arithmetic, as used here.
- elementary science.** Such subjects as nature study, hygiene, and general science, as used here.
- emotional adjustment inventory.** See *adjustment inventory*.
- equated scores.** Derived scores that are comparable from test to test of a certain battery.
- equivalent forms.** Duplicate or equal forms of a standardized test that yield closely similar scores.
- error of grouping.** A variable error introduced by the practice of combining in class intervals scores or measures that are unlike.
- essay examination.** A test to which the pupil ordinarily responds with written discussion of issues raised in several broad questions.
- evaluate.** To test, measure, and appraise the "whole" child by the use of tests and a wide variety of non-test tools and techniques.
- examination.** See *test*.
- exercise.** A unit of a test governed by a specific set of directions.
- expectancy.** The standard of future achievement held reasonable for the individual pupil.

**expressive language arts.** Language and grammar, handwriting, and spelling, as used here.

**extensive sampling.** See *adequacy*.

**extrapolation.** The process of locating a point beyond two or more known points in accordance with the conditions operating in the given case.

**factor analysis.** A method widely used in the study of the nature of mental and other abilities.

**factored test.** A test from the use of which several scores representing different factors of general ability are obtained.

**faculty theory.** The theory that intelligence consists of a large number of relatively independent and largely correlated and specialized abilities, such as memory and imagination.

**feeble-minded.** The term used to designate persons of inferior intelligence having IQs below 70.

**fine arts.** Music and art, as used here.

**first quartile ( $Q_1$ ).** The point on a scale of values below which 25 per cent of the cases fall; the 25th percentile.

**"footrule" coefficient.** An index giving an estimate of test reliability useful when only one form of a test is available.

**form.** One of the two or more arrangements of closely similar or equivalent standardized tests that in itself constitutes a testing unit.

**frequency ( $f$ ).** The number of measures in a given class interval of a frequency distribution.

**frequency curve.** See *frequency polygon*.

**frequency distribution.** The table in which scores or other measures are classified.

**frequency polygon.** A type of graphical representation used to show the manner in which scores in a frequency distribution are distributed.

**fulcrum.** The axis upon which a lever is supported and rotated.

**general ability.** Closely similar to general intelligence; ability to learn.

**general achievement test.** An educational test covering several fields of study and ordinarily adapted for use in several grades.

**general intelligence test.** A test of general mental ability.

**genius.** A person of superior intelligence having an IQ of 140 or above.

**gradation.** See *classification*.

**grade.** The administrative division of the school that indicates the educational level of the pupil.



- grade equivalent.** The score derived from grade norms on a standardized test.
- grade norms.** Tables of values representing typical or average performance on standardized tests for pupils in different grades.
- group dynamics.** Interactions among the individual members of a group engaged in some cooperative activity.
- group factors of intelligence.** The different phases or aspects of intelligence resulting from scientific analyses of intellectual abilities.
- group-factor test.** A type of intelligence test from the use of which separate scores for several aspects of mental ability are obtained.
- group test.** A test that can be administered to a number of pupils at the same time.
- grouping.** The process of classifying and tabulating data into class intervals or steps.
- half-sum.** A term used in the calculation of the median.
- halo effect.** The tendency of a teacher to be influenced in rating pupil performance by impressions previously acquired.
- health education.** Health facts, attitudes, and practices, as used here.
- histogram.** A type of graphical representation employing only horizontal and vertical lines.
- identification test.** A test of ability to recognize and name objects shown or pictured.
- idiot.** A feeble-minded person having an IQ below 25.
- imbecile.** A feeble-minded person having an IQ from 25 to 49.
- index of brightness (IB).** A measure of brightness somewhat similar to the intelligence quotient in meaning.
- index of studiousness.** The difference between a pupil's rank in his class on intelligence and on achievement.
- individual differences.** The observed or measured variation of individuals in ability, progress, or achievement.
- individual test.** A test that can be administered to only one pupil at a time.
- informal objective test.** A teacher-made objective test.
- instructional objective.** An aim or purpose of instruction.
- instructional outcome.** A result of instruction stated in terms of pupil behavior.
- instructional test.** A test used directly in teaching a unit of material.
- integral limits.** The lower and upper whole-number limits of a class interval in a grouped frequency distribution.

**intelligence.** The ability to adapt oneself to changing conditions; ability or power to learn.

**intelligence quotient (IQ).** The ratio between mental age and chronological age.

**intelligence test.** A test that measures ability to learn or to profit from experience.

**intensive sampling.** A narrow and inadequate selection of test items that results in a test of too little scope or range.

**interests.** An instructional or learning outcome represented by a mental set that urges a person to act in a certain manner.

**interests inventory.** An instrument used in the determination of pupil interests in various fields or areas of performance.

**interpolation.** The process of locating an intermediate point between two known points in accordance with the conditions operating in the given case.

**interpretive test.** An achievement test in which items are based on data presented in verbal, numerical, or graphical form.

**interval.** See *class interval*.

**interval deviation.** The number of class-interval units by which a certain interval in a frequency distribution differs from the interval in which the arithmetic mean is assumed to lie.

**interview.** A personal conference technique frequently used in diagnosis and in the evaluation of attitudes.

**inventory.** A personal-report type of scale or test commonly used in measuring personality.

**inventory test.** A test used as a preliminary check on the degree of mastery existing prior to instruction.

**item count.** A method used to determine whether test items properly discriminate between pupils of various ability levels.

**job analysis.** The process of breaking down a certain task into its elements or component parts.

**knowledges.** An instructional or learning outcome represented by the ability to recall or recognize facts, persons, places, or things.

**learning outcome.** A result of experience in or outside of the school stated in terms of pupil behavior.

**listening test.** A test of ability to comprehend spoken language.

**logical validity.** See *psychological validity*.

**machine-scored test.** A test that can be scored by the use of an electrical or mechanical scoring machine.



- manuscript writing.** A free-hand style of lettering in which the letters are not connected as in common script writing.
- mark.** The teacher's numerical or letter evaluation of pupil achievement in a course or area of performance.
- mastery test.** An achievement test designed to determine how thoroughly a pupil has learned certain facts or skills.
- matching exercise.** A type of test exercise to which the pupil responds by attempting to pair the related items in two or more columns of related facts or ideas.
- mathematics.** See *elementary mathematics*.
- mean.** See *arithmetic mean*.
- measure.** To test by means of standardized and teacher-made instruments mainly in the fields of achievement and intelligence; a test score or other numerical rating.
- median (Mdn.).** The point on the scale below which half of the measures in a frequency distribution fall.
- mental ability.** Ability or power to learn; nearly synonymous with intelligence.
- mental age (MA).** The intelligence or mental ability of a person expressed in terms of the chronological age of which his mental ability is typical.
- mental test.** A test of intelligence or personality, as distinguished from an educational test.
- metronoscope.** A device for exposing strips of reading material for reading drill.
- midmeasure.** The middle measure of a series of values arranged in order of magnitude.
- midpoint.** The exact middle of a class interval in a frequency distribution.
- moron.** A feeble-minded person having an IQ from 50 to 69.
- multiple-choice item.** A type of test item to which the pupil responds by attempting to select the correct or best response from the several alternatives given.
- multiple-factor test.** See *group-factor test*.
- multiple-response item.** A type of test item to which the pupil responds by attempting to indicate all correct answers.
- new-type examination.** See *informal objective test*.
- non-language test.** A test not involving the use of words in its administration or taking, e.g., a test given by pantomime.
- non-test tool.** An instrument other than a test used in measuring pupil behavior.

**non-verbal test.** A test not involving the use of words by the pupils in attaching meaning to the items, e.g., a figure analogies test.

**normal.** Typical in progress, growth, development, or distribution.

**normal curve.** The graphic representation of a large number of cases in the selection of which chance was operative.

**norms.** The median or average performances on standardized tests of pupils of different ages or grade placement or of school groups.

**object test.** A test involving the use of three-dimensional objects.

**objective.** An aim or purpose.

**objective test.** A test for which the scoring procedure eliminates subjective opinion and judgment.

**objectivity.** An examination criterion indicating the degree to which subjective opinion and judgment are eliminated in the process of scoring it.

**objectivity coefficient.** A correlation coefficient used in describing the objectivity of a test.

**observational methods.** Certain techniques of personality study, e.g., directed observation and the anecdotal method.

**ogive.** See *cumulative frequency graph*.

**ophthalmograph.** A binocular camera used in measuring eye movements during reading.

**oral examination.** A test administered and answered orally.

**outcome.** A result stated in terms of pupil behavior.

**percentile.** One of the ninety-nine points that divide a distribution into one hundred equal areas.

**percentile curve.** See *cumulative frequency graph*.

**percentile-grade norms.** Tables of percentile ranks on test scores for pupils in different school grades.

**percentile norms.** Tables of values representing percentile ranks of scores on standardized tests for certain subjects or certain grades.

**percentile rank.** The position assigned to a score in an array for which the scores are divided into one hundred equal divisions in descending order.

**performance.** The accomplishment, achievement, or behavior of the pupil.

**performance test.** A test to which the pupil typically responds by motor or manual rather than by verbal behavior.

**personal constant (PC).** A measure of brightness obtained by the use of Heinis growth units for both the mental age and the chronological age.



**personal reports.** The responses given by the pupil on certain types of personality scales and inventories.

**personality.** An individual's total behavior in social situations.

**personality inventory.** An instrument that measures such intangible aspects of behavior as attitudes, interests, and adjustment.

**personality quotient (PQ).** A quotient sometimes used in the measurement of total personality.

**physical education.** Motor skills, attitudes, and activities, as used here.

**point score.** See *raw score*.

**power test.** A test that measures the difficulty of the task the pupil is just able to perform.

**practicality.** An examination criterion indicating the degree to which a test possesses certain utilitarian characteristics.

**practice effect.** The influence of a previous experience with a test on a later encounter with the same or a similar test.

**practice exercise.** A few sample items preceding a test designed to familiarize the pupils with the nature of the test.

**practice test.** See *drill test*.

**preference.** A liking for or predisposition toward some person, activity, or practice.

**preventive teaching.** Steps taken at the time of initial instruction to guard against the later appearance of defects or difficulties in pupil learning.

**primary mental abilities.** A term often applied to factors of mental ability.

**product scale.** See *source scale*.

**profile chart.** A device used for graphical representation of scores made by the pupil on the various parts of certain achievement, intelligence, and personality tests.

**prognostic test.** A test used to predict future success in specific subjects or fields.

**progress record.** A device similar to a profile chart on which pupil progress from year to year can be shown graphically for certain achievement tests.

**projective method.** A technique of personality study involving the observation of how a person reacts to certain toys and materials.

**prophecy formula.** The Spearman-Brown formula used in estimating test reliability from a correlation coefficient between scores on "chance-halves" of a test.

**psychological examination.** See *intelligence test*.

**psychological validity.** Evidence of test validity resulting from a logical dissection of a total learning process.

- quality scale.** A series of standard graded samples with which the production of the pupil is compared in evaluating performance in such areas as handwriting and composition.
- quartile.** One of the three points that divide a distribution into four equal areas.
- quiz.** A short achievement test covering an assignment or a restricted unit of course content.
- quotient.** A ratio designed to reveal in a single numerical index the relative position of the pupil on two related variables.
- range (R).** The distance from the lowest to the highest score in a series of scores.
- rate score.** A score expressing a pupil's rate of work.
- rate test.** A test that measures speed of performance on tasks of uniform difficulty.
- rating scale.** An instrument used by a teacher or other person in the evaluation of pupil personality or achievement.
- raw score.** The quantitative result obtained directly from the scoring of a test or scale.
- readiness test.** A test that measures the ability of the pupil to undertake a new type of specific learning.
- real limits.** The actual or true lower and upper limits of a class interval in a frequency distribution.
- recall item.** A type of test item to which the pupil responds by writing words, numbers, or phrases to complete the meaning of a statement.
- receptive language arts.** Reading and study methods, as used here.
- recognition item.** A type of test item to which the pupil responds by indicating the truth or falsity of statements, selecting the correct or best answer from among several given, or indicating the proper pairing of related items.
- relative rank.** The position assigned to a score in an array for which the scores are arranged in descending order.
- reliability.** An examination criterion indicating the degree to which a test measures what it does measure; consistency of measurement.
- reliability coefficient.** The correlation coefficient obtained between scores made by the same pupils on two equivalent forms of a test.
- remedial.** Having as a purpose the correction of observed difficulties and weaknesses in performance.
- remediation.** See *corrective teaching*.
- retesting coefficient.** An estimate of test reliability that can be obtained when only one form of a test is available.



- sampling.** The process of selecting a limited number of cases or items that will be representative of the large group from which they are chosen.
- scale.** An instrument used by the scorer in evaluating pupil performance or by the test-maker in constructing a test; the continuum from the lowest to the highest score in a frequency distribution.
- scaled score.** A derived score based upon deviation from the arithmetic mean in units of one-tenth of a standard deviation for a group established in a certain manner.
- scaled test.** A test in which the items are arranged in an order of increasing difficulty.
- school norms.** Tables of percentile ranks based on the mean test scores of pupils in different schools.
- science.** See *elementary science*.
- scorability.** An examination criterion indicating the characteristics of a test that make for ease and simplicity in scoring it.
- score.** A quantitative description of performance.
- score card.** A short and simple type of rating scale used in the evaluation of products made by pupils.
- score deviation.** The number of score units by which a certain score in a frequency distribution differs from the mean or the assumed mean.
- self-marking test.** A test that does not require the use of scoring keys or machines in the scoring process.
- sigma ( $\sigma$ ).** See *standard deviation*.
- simple recall item.** A type of test item to which the pupil responds by writing the word, number, or phrase that he believes will correctly complete a statement or answer a question.
- simulated-conditions test.** A test in which the conditions represent or approximate in nature those of the ultimate performance it is used to evaluate.
- skills.** An instructional or learning outcome involving some form of physical or motor performance.
- social studies.** Such content subjects as history, civics and government, and geography, as used here.
- social utility.** A point of view basic to the selection of curricular materials which holds that subject matter should contribute definitely to child and adult needs.
- sociogram.** A graphic device for representing interpersonal relations within a group of pupils.
- sociometric methods.** Certain procedures adapted from sociology for use in evaluating pupil behavior.

- source scale.** A series of items of graded difficulty from which tests can be constructed, e.g., a spelling scale.
- specific determiners.** Characteristics of true-false test items that seem to determine in part the nature of the correct response.
- speed test.** See *rate test*.
- standard.** A level of performance agreed upon by experts or established by local school officers as a goal of pupil attainment.
- standard deviation (S.D.).** The most widely useful measure of variability or dispersion.
- standard error of measurement.** A measure of score accuracy used in estimating test reliability.
- standard score.** A derived score based upon deviation from the arithmetic mean in terms of the standard deviation.
- standardization.** The process of constructing a test and establishing norms for it.
- standardized test.** A test for which the items have been carefully selected and evaluated and which is accompanied by norms.
- statistical validity.** Evidence of test validity shown by correlational relationship or other statistical procedures.
- step.** See *class interval*.
- structured inventory.** A personal-report type of personality scale to which the pupil must respond in one of the several prescribed ways.
- subjectivity.** The degree to which measurement results are influenced by personal opinions or judgment.
- sub-total.** A term used in the calculation of the median.
- survey test.** A test that measures general achievement in certain subjects or fields.
- synthesis.** The process of combining underlying and somewhat isolated skills so that they form an effective unit.
- T-score.** A derived score based upon deviation from the arithmetic mean in units of one-tenth of a standard deviation.
- tabulation.** The process of grouping and classifying data; the distribution into which data are classified.
- tachistoscope.** A device for exposing strips of reading material for reading drill.
- talent.** See *aptitude*.
- taste.** See *preference*.
- teacher-made test.** A test constructed by the teacher, such as the essay and informal objective tests.
- teacher's mark.** See *mark*.
- technique.** A procedure or method.



- telebinocular.** A type of stereoscope adjustable for various distances.
- test.** In the general sense any instrument used in the measurement of any educational or mental ability and in a specific sense an instrument used by the pupil and ordinarily involving the use of paper and pencil; to measure by the use of tests.
- test battery.** A group of several tests covering a number of different subjects and intended for use in testing over wide areas.
- test item.** The smallest unit of a test.
- test rating scale.** A scale used in the evaluation of tests for specific uses.
- third quartile ( $Q_3$ ).** The point on a scale of values below which 75 per cent of the cases fall; the 75th percentile.
- time-limit test.** A test on which the working time allowed pupils is rigidly prescribed.
- tool.** An instrument of a test or non-test type used in measuring pupil behavior.
- tool subjects.** Fields in which achievement consists mainly in the acquisition of skills and techniques useful in further learning, as reading, arithmetic, and spelling.
- traditional examination.** See *essay examination*.
- true-false item.** A type of alternate-response item to which the pupil responds by indicating whether a statement is true or false.
- two-factor theory.** The theory that intelligence consists of a general factor, many specific factors, and a number of group factors.
- understandings.** An instructional or learning outcome involving comprehension of meaning and of the uses and significance of what has been learned.
- unstructured technique.** A projective method of personality measurement in which the pupil has wide freedom in his manner of responding.
- utility.** An examination criterion indicating the degree to which a test serves a definite need.
- validity.** An examination criterion indicating the degree to which a test measures what it purports to measure.
- validity coefficient.** A correlation coefficient used in expressing the validity of a test.
- variability.** The spread or dispersion of scores.
- variable.** A quality that may exist in different amounts.
- variable error.** A type of deviation from complete accuracy that results from the tendency of persons to vary in their judgments from time to time.

**verbal test.** A test involving the use of language in the form of words by the pupil in attaching meaning to, responding to, or both attaching meaning to and responding to the items.

**work-limit test.** A test on which sufficient time is allowed for all or nearly all pupils to complete their work.

**work-sample test.** A test consisting of a representative portion of the ultimate performance it is used to evaluate.

**work-type reading.** The types of silent reading skills commonly utilized in study.

**yes-no item.** A type of alternate-response item to which the pupil responds by an affirmative or negative answer to a question.



## ***Appendix***

- Acorn Publishing Co., Inc., Rockville Centre, N. Y.  
American Council on Education, 1785 Massachusetts Ave., Washington 6, D. C.  
Association Press, 347 Madison Ave., New York, N. Y.  
Avent, Jos. E., Box 1455, Knoxville, Tenn.  
Bureau of Educational Measurements, Kansas State Teachers College, Emporia, Kan.  
Bureau of Educational Research and Service, University of Iowa, Iowa City, Iowa.  
Bureau of Publications, Teachers College, Columbia University, New York 27, N. Y.  
California Test Bureau, 5916 Hollywood Blvd., Los Angeles 28, Cal.  
Cooperative Test Service (See Educational Testing Service)  
Educational Test Bureau, Inc., 721 Washington Ave., S. E., Minneapolis, Minn.  
Educational Testing Service, 20 Nassau St., Princeton, N. J.  
Laidlaw Brothers, Inc., 320 East 21st St., Chicago, Ill.  
Personnel Press, Inc., 188 Nassau St., Princeton, N. J.  
Psychological Corporation, 522 Fifth Ave., New York 18, N. Y.  
Public School Publishing Co., 509-13 North East St., Bloomington, Ill.  
Science Research Associates, Inc., 57 West Grand Ave., Chicago 10, Ill.  
Scott, Foresman and Co., 623 S. Wabash Ave., Chicago 5, Ill.

Southern California School Book Depository (See California Test Bureau)

Stanford University Press, Stanford, Cal.

State University of Iowa, Iowa City, Iowa.

Steck Co., Austin 1, Texas.

C. H. Stoelting Co., 424 N. Homan Ave., Chicago 24, Ill.

University of Minnesota Press, Minneapolis 14, Minn.

World Book Co., 313 Park Hill Ave., Yonkers 5, N. Y.



## *Index of Names*

- Adams, Georgia S., 182, 472, 513, 550,  
     578, 580  
 Adams, Mary A., 476, 480  
 Adkins, Dorothy C., 196, 209, 215  
 Aldington, Richard, 279  
 Allen, Richard D., 180, 186, 227, 474,  
     475, 569  
 Allen, Wendell C., 236  
 Anastasi, Anne, 28, 29, 32, 35  
 Anderson, Gladys L., 306  
 Anderson, Harold H., 306  
 Anderson, Howard R., 466, 471, 478, 479  
 Anderson, Irving H., 416  
 Anderson, Rose G., 249, 263  
 Anderson, W. N., 440, 441  
 Apgar, Sara L., 562  
 Arnold, Dwight L., 524  
 Arny, Clara B., 203, 212, 213, 215  
 Ashbaugh, Ernest J., 51, 442, 459  
 Ashbrook, Willard P., 556, 559, 560, 561  
 Ashburn, Robert, 158  
 Ayer, F. C., 527  
 Ayres, Leonard P., 24, 35, 440, 450, 453  
  
 Baker, Harry J., 295  
 Bales, Robert F., 302  
 Ballenger, H. L., 97, 422, 434  
 Barnes, Elinor J., 158  
 Barnes, Melvin W., 543  
 Barnett, Sidney W., 479  
 Beach, Frank A., 535  
  
 Bell, Hugh A., 60  
 Bell, John E., 62, 298, 306  
 Bender, William, Jr., 158  
 Bennett, George K., 255  
 Berg, Harry D., 184, 473  
 Berman, Louis, 279  
 Bettelheim, Bruno, 544  
 Betts, Emmett A., 401, 403, 414, 416,  
     459  
 Binet, Alfred, 29, 31, 260  
 Bingham, Walter V., 62, 236, 275  
 Bixler, H. H., 442  
 Blanton, Smiley, 426  
 Block, Jean F., 307  
 Bloom, Samuel L., 215  
 Bond, Eva, 416  
 Bond, Guy L., 416  
 Bovard, John F., 561  
 Boyd, William, 21  
 Boynton, Holmes, 503  
 Boynton, Marcia, 459  
 Boynton, Paul L., 63, 243, 276  
 Brace, David K., 561  
 Branom, M. E., 570  
 Bregman, Elsie O., 260  
 Brody, William, 158  
 Broer, Marion R., 562  
 Brooks, B. Marian, 528, 543  
 Brooks, Fowler D., 540  
 Broom, M. E., 18, 135, 196, 275, 369,  
     416, 479, 502, 543  
 Brouwer, Paul J., 224

- Brown, Clara M., 209  
 Brown, Harry A., 528, 543  
 Brown, James I., 393, 399, 408, 416  
 Brownell, William A., 84, 196, 483, 484, 486, 502  
 Brueckner, Leo J., 502  
 Buckingham, Guy E., 524  
 Buros, Oscar K., 84, 124, 135, 236, 275, 276, 306, 416, 459, 479, 502, 524, 543, 561, 580  
 Burt, Cyril L., 135  
 Buswell, Guy T., 411, 502  
 Burton, William H., 445  
 Butterfield, Marguerite, 445, 448, 461  
  
 Cain, Maud, 479  
 Caldwell, Otis W., 22, 23, 140  
 Calkins, Mary W., 158  
 Callewaert, H., 459  
 Campbell, Doak S., 465, 480  
 Carlsen, G. R., 408  
 Carpenter, Aileen, 562  
 Carr, Edwin R., 462, 479  
 Carter, R. E., 151, 152, 159  
 Case, Adelaide T., 286  
 Cason, Hulsey, 158  
 Cattell, J. McKeen, 29  
 Cattell, Psyche, 263  
 Cattell, Raymond B., 62, 243, 306  
 Chadwick, E. B., 23  
 Chaille, Dr. E. S., 29  
 Chall, Jeanne S., 402  
 Chapman, J. Crosby, 268  
 Chave, E. J., 64, 285  
 Chenoweth, Laurence B., 560  
 Clark, Willis W., 108, 208, 226, 254, 365, 473, 576, 581  
 Clarke, H. Harrison, 562  
 Conard, Edith U., 454  
 Conrad, Herbert S., 136, 196, 369  
 Cook, Walter W., 62, 136, 196, 442, 459  
 Cooper, Bernice I., 562  
 Courtis, Stuart A., 22, 23, 140  
 Cowell, Charles C., 562  
 Cozens, Frederick W., 556, 559, 560, 561, 563  
 Cram, Fred D., 433  
 Crawford, John R., 158, 197, 338, 369, 390  
 Cronbach, Lee J., 62, 276  
 Crow, Lester D., 178  
 Cubberley, Hazel J., 560  
  
 Cureton, Edward E., 84  
 Cureton, Thomas K., 562  
 Curtis, Francis D., 524  
  
 Dale, Edgar, 402  
 Darley, John G., 136, 236, 276, 306, 338, 369, 390  
 Davis, Allison, 261  
 Davis, Frederick B., 136  
 Davis, Ira C., 520  
 Davis, Robert A., 158  
 Davis, Warren M., 518  
 Dawson, Mildred, 459  
 Dean, Charles D., 543  
 Dearborn, Walter F., 411, 416  
 DeGraff, M. H., 472  
 Denny, E. C., 474  
 Derryberry, Mayhew, 551  
 Dewey, John, 493  
 Diederich, Paul B., 459, 544  
 Dolch, Edward W., 416, 459  
 Doolittle, N. A., 18  
 Dougherty, Mary L., 459  
 Douglass, Harl R., 169, 267  
 Downing, Elliot R., 524  
 Drake, Raleigh M., 532  
 Driscoll, Gertrude, 236  
 Dunkel, Harold B., 306, 543  
 Dunlap, Jack W., 134  
 Durrell, Donald D., 408, 414, 416  
 Dyer, Joanna T., 562  
 Dykema, Karl W., 123  
 Dykema, Peter W., 527  
  
 Ebel, Robert L., 187, 196, 235  
 Edgeworth, F. Y., 141  
 Eells, Kenneth, 261  
 Eells, Walter C., 144  
 Elliott, Edward C., 141, 142, 143, 144  
 Ely, Lena A., 472  
 Engelhart, Max D., 62, 84, 158, 197, 236  
 Espenschade, Anna, 556, 559, 560, 561  
  
 Farnsworth, Paul R., 543  
 Faulkner, Ray N., 543, 544  
 Ferguson, Leonard W., 306  
 Finch, F. H., 416  
 Findley, Warren G., 136, 236  
 Fisher, Rev. George, 23-24  
 Fitzgerald, J. A., 440  
 Flanagan, John C., 84, 92, 136, 369



- Flesch, Rudolf, 402  
 Foley, John P., Jr., 28, 29, 32, 35  
 Foran, Thomas G., 459  
 Forsyth, Elaine, 184, 466, 471, 473, 479  
 Frank, Lawrence K., 306  
 Franzen, Raymond, 551  
 Frederickson, Norman, 216  
 Freeman, Frank N., 29, 30, 35, 62, 239,  
     240, 262, 263, 264, 276, 306, 450,  
     454, 459, 461  
 Freeman, Frank S., 236, 276, 306  
 French, Esther L., 562, 563  
 Friedberg, Jean, 307  
 Friedman, Bertha S., 276  
 Froehlich, Clifford P., 136, 236, 276, 306,  
     338, 390  
 Frutchey, Fred P., 524  
 Fryer, Douglas, 63, 287, 306
- Gage, N. L., 18, 136, 159, 197, 216, 277,  
     338, 370, 390  
 Galton, Francis, 28, 32, 282  
 Garrett, Henry E., 21, 28, 35, 279, 338,  
     369, 390  
 Gates, Arthur I., 402, 410, 412, 414, 416,  
     417, 549, 571, 580  
 Gerberich, J. Raymond, 35, 64, 84, 85,  
     140, 158, 168, 177, 197, 198, 237,  
     369, 390, 417  
 Gibson, James J., 215  
 Gildersleeve, Glenn, 535, 544  
 Gillenwater, V. W., 416  
 Gilmore, John V., 406  
 Glaser, Edward M., 222  
 Glassow, Ruth B., 562  
 Glennon, Vincent J., 502  
 Goddard, H. H., 30, 245  
 Good, Carter V., 63  
 Goodenough, Florence L., 29, 35, 236,  
     260, 276, 306  
 Grant, Parks, 544  
 Graves, Maitland, 544  
 Gray, Albert, 499  
 Gray, Hob, 110, 571  
 Gray, William S., 395, 400, 402, 407,  
     411, 412, 414, 417, 422, 460  
 Greene, Charles E., 502  
 Greene, Edward B., 63, 84, 197, 216,  
     276, 307, 338, 369  
 Greene, Harry A., 49, 97, 158, 197, 210,  
     214, 216, 251, 338, 369, 390, 395,  
     417, 422, 433, 434, 459, 460, 472
- Griffiths, Nellie L., 252  
 Grossnickle, Foster E., 502  
 Gudakunst, Don W., 562  
 Guilford, J. P., 84
- Haefner, Ralph, 18, 137, 277, 418, 461  
 Haggerty, M. E., 295  
 Hamalainen, Arthur E., 479  
 Hanna, Lavone A., 63, 219  
 Harris, Albert J., 400, 402, 417  
 Harris, Chester W., 460, 544  
 Harrison, M. Lucille, 404, 417  
 Harsh, C. M., 307  
 Hartshorne, Hugh, 32, 307  
 Hartung, Maurice L., 236  
 Havighurst, Robert J., 261, 307  
 Hawkes, Herbert E., 66  
 Hayes, Margaret, 296  
 Heath, Marjorie L., 562  
 Heffernan, Helen W., 460  
 Heil, Louis M., 515, 524  
 Heinis, H., 263  
 Hendrickson, Gordon, 544  
 Henri, V., 29  
 Herrick, John H., 479  
 Herring, John P., 245, 579, 581  
 Hilden, Arnold H., 263  
 Hildreth, Gertrude H., 136, 233, 252,  
     460, 490, 560, 580  
 Hilpert, Robert S., 537  
 Hippocrates, 279  
 Holzinger, Karl J., 276  
 Horn, Ernest, 440, 441, 443, 460, 484  
 Howell, Hazel W., 462-63  
 Huey, E. B., 411  
 Huffaker, C. L., 267  
 Hull, Clark L., 63, 276  
 Humphreys, Lloyd G., 63, 276  
 Hunt, Thelma, 35, 63
- Jacobs, Robert, 307  
 Jennings, Helen H., 301  
 Johnson, Granville B., 556  
 Johnson, Leslie W., 460  
 Johnson, Philip G., 524  
 Johnson, W. P., 498  
 Johnson, Wendell, 460  
 Jordan, A. M., 18, 84, 158, 197, 276,  
     307, 417, 460, 479, 502, 524, 544,  
     563, 580  
 Jung, C. G., 32, 279

- Kandel, I. L., 141, 158  
 Karnes, M. Ray, 18, 63, 84, 197, 216,  
 338, 370  
 Kelley, Truman L., 92, 243, 479, 513,  
 568  
 Kelly, Fred J., 153  
 King, Edith, 472  
 Kirby, C. Valentine, 537, 538, 544  
 Klar, Walter H., 537, 544  
 Knight, Edgar W., 21, 22  
 Knight, Frederick B., 49, 483, 502  
 Kohn, Clyde F., 479  
 Koos, L. V., 236, 450  
 Kopel, David, 418  
 Kraepelin, Emil, 32  
 Kretschmer, Ernst, 279  
 Krey, August C., 479  
 Kuder, G. F., 74, 387  
 Kuhlmann, F., 30, 245, 249, 263  
 Kwalwasser, Jacob, 527, 534, 535, 536,  
 544  
  
 Lang, Albert R., 22  
 Lange, Elizabeth, 563  
 LaPorte, William R., 554  
 Lawson, D. E., 158  
 Leary, Bernice E., 402, 548  
 Lee, J. Murray, 18, 26, 63, 136, 174, 197,  
 338, 369  
 Lee, Richard E., 524  
 Lessenger, W. E., 267  
 Lester, John A., 417, 461  
 Lewerenz, A. S., 401  
 Lewis, Don, 531  
 Lewis, Hugh B., 236  
 Lewry, Marion E., 461  
 Limbert, Paul M., 286  
 Lincoln, Edward A., 63, 370  
 Lindquist, E. F., 18, 66, 73, 74, 84, 172,  
 197, 335, 338, 370, 390, 417, 461,  
 478, 490, 491, 575  
 Lorge, Irving, 402  
 LuPone, O. J., 516, 524  
  
 Maller, Julius B., 307  
 Mann, C. R., 66  
 Mann, Horace, 22, 23, 140  
 Manuel, Herschel T., 527  
 Marckwardt, Albert H., 461  
 Martin, Lureata R., 545  
 Martin, W. A. P., 20  
 Maurer, Katharine M., 260  
 Maxfield, Francis N., 264  
 May, Mark A., 32, 307  
 McBroom, Maude, 422  
 McCall, William A., 18, 26, 303, 304,  
 305, 551, 579, 580, 581  
 McCauley, Clara J., 534  
 McCloy, Charles H., 555, 556, 557, 560,  
 563  
 McConn, Max, 18, 136  
 McCune, George H., 418  
 McGuire, Christine, 237  
 McKee, Paul, 417, 421, 422, 461  
 McNamara, Walter J., 64, 198, 338, 370,  
 391, 461, 480, 525  
 McNemar, Quinn, 276  
 McSwain, E. T., 237  
 Meeker, Marchia, 261  
 Meeker, Ronald W., 544  
 Meier, Norman C., 527, 541, 544  
 Merrill, Maud A., 63, 245, 258, 264, 265,  
 277  
 Meyer, George, 159  
 Michaelis, John U., 479  
 Micheels, William J., 18, 63, 84, 197,  
 216, 338, 370  
 Moffatt, Maurice P., 462, 463, 479  
 Monroe, Marion, 417, 418  
 Monroe, Walter S., 33, 35, 151, 152,  
 159  
 Mooney, Ross L., 294  
 Moore, Bruce V., 236  
 Moore, J. E., 545  
 Morrison, Whitelaw R., 560  
 Morse, Horace T., 418, 466, 471, 479  
 Mosier, Charles I., 197  
 Mulgrave, Dorothy I., 426, 461  
 Munro, Thomas, 544  
 Münsterberg, Hugo, 31  
 Murphy, Helen A., 408  
 Mursell, James L., 276, 307, 544  
 Myers, M. Claire, 197  
  
 Neilson, Neils P., 560, 563  
 Nelson, M. J., 18, 63, 84, 136, 197, 276,  
 338, 370, 390, 474, 503, 545, 581  
 Newkirk, Louis V., 210, 214, 216  
 Nichols, Ralph G., 399  
 Noll, Victor H., 505, 506, 519, 525  
 Nyswander, Dorothy, 563



Odell, Charles W., 25, 152, 159, 197, 338,  
370, 390  
Ojemann, Ralph H., 401  
Olson, Willard C., 63, 280, 295, 307  
Orata, Pedro T., 480  
Orleans, Jacob S., 18, 63, 136, 276  
O'Shea, M. V., 440  
OSS Assessment Staff, 236  
Otis, Arthur S., 30, 248, 264, 579  
O'Toole, Charles E., 47, 203  
Overman, James R., 483

Parry, Margaret E., 503  
Paterson, Donald G., 256  
Pearson, Karl, 28  
Peters, Charles C., 84, 390  
Peterson, Joseph, 28  
Pintner, Rudolf, 35, 56, 63, 256, 293, 367  
Piper, A. H., 251  
Plato, 21  
Popenoe, Herbert E., 512, 514  
Powell, Norman J., 158  
Powers, Samuel Ralph, 507, 515, 525  
Pressey, S. L., 60, 158  
Price, Helen G., 197  
Pritchard, Maralyn W., 33, 36

Quillen, I. James, 63, 219  
Quintillian, 21

Rankin, Paul T., 393  
Raths, Louis, 236  
Read, John G., 184, 513, 525  
Reavis, William C., 33  
Reiner, William B., 525  
Remmers, H. H., 18, 63, 136, 159, 197,  
216, 277, 285, 338, 370, 390  
Rice, Dr. J. M., 24  
Richardson, M. W., 74, 262, 370, 387  
Rinsland, Henry D., 197  
Robertson, D. A., 297  
Rodgers, Elizabeth G., 563  
Rogers, Carl R., 292, 293  
Rogers, Frederick R., 555, 563  
Rogers, J. Lloyd, 110, 571  
Roos, Frank J., 545  
Ross, C. C., 18, 35, 84, 136, 159, 198,  
338, 370, 390  
Ruch, Giles M., 26, 49, 144, 237, 472,  
512, 514, 535, 536  
Rugen, Mabel E., 563

Russell, Charles, 21  
Russell, David H., 418  
Russell, Naomi, 563  
Ryans, David G., 216

Saetveit, Joseph, 531  
Sargent, Helen, 33, 35, 307  
Sauble, Irene, 503  
Scates, Douglas E., 35, 85  
Scheidemann, Norma V., 20  
Schoen, Max, 527, 545  
Schrickel, H. G., 307  
Schroeder, E. M., 563  
Schultz, Harold A., 545  
Schurig, Jennie C., 562  
Scott, M. Gladys, 563  
Seashore, Carl E., 527, 531, 545  
Seashore, Harold G., 255  
Segel, David, 26, 35, 122, 136, 237  
Sexson, John A., 182, 472, 513, 550,  
578, 580  
Shaffer, Laurance F., 279, 280  
Shane, Harold G., 237  
Shepard, Lon A., 433  
Sheviakov, George V., 307  
Shotwell, Anna M., 18, 36, 64, 198, 277,  
338, 391, 418, 461, 480, 503, 525,  
545, 581  
Silance, E. B., 63  
Simon, Théodore, 29  
Sims, Verner M., 154, 159, 299  
Siro, Einar E., 216  
Skeels, Harold M., 260  
Smith, Allan B., 136  
Smith, Dora V., 418  
Smith, Eugene R., 27, 370  
Smith, Samuel, 178, 573  
Socrates, 20  
Sommer, R., 32  
Soper, Wayne, 535  
Spache, George, 418, 461, 503  
Spaulding, Geraldine, 85, 198  
Spearman, Charles, 31, 241  
Speer, Robert K., 178, 573  
Spitzer, Herbert F., 39, 169, 484, 503  
Squire, Russel N., 545  
Stagner, Ross, 63, 307  
Stalnaker, John M., 144, 147, 153, 154,  
159  
Stanton, Hazel M., 527, 545  
Starch, Daniel, 141, 142, 143, 144  
Stoddard, George D., 240, 260, 277

- Stone, Cliff W., 25  
 Strang, Ruth, 547, 548, 549, 563  
 Stratemeyer, Clara G., 544  
 Strickland, Ruth G., 461  
 Strong, Edward K., Jr., 288, 289, 290  
 Stroud, James B., 404  
 Studebaker, J. W., 49  
 Suelz, Ben A., 503  
 Sullivan, Elizabeth T., 254  
 Super, Donald E., 216, 277, 307  
 Symonds, Percival M., 63, 174, 237, 260, 268, 307  
  
 Taba, Hilda, 237, 307, 370  
 Terman, Lewis M., 30, 63, 245, 248, 257, 258, 264, 265, 277  
 Theophrastus, 279  
 Thomson, Godfrey H., 31  
 Thorndike, Edward L., 24, 25, 260, 418, 440, 441, 461, 527  
 Thorndike, Robert L., 85  
 Thurstone, L. L., 63, 64, 241, 253, 277, 285  
 Thut, I. N., 64, 85, 140, 168, 198, 237  
 Tidyman, Willard F., 445, 448, 461  
 Tiegs, Ernest W., 108, 208, 226, 254, 365, 370, 461, 473, 576, 581  
 Todd, Jessie M., 545  
 Trabue, M. R., 422  
 Travers, Robert M. W., 85, 198  
 Travis, Lee E., 427, 428, 461  
 Traxler, Arthur E., 18, 85, 137, 159, 198, 237, 277, 280, 307, 370, 418, 461  
 Trimble, Otis C., 159  
 Tyler, Ralph W., 26, 27, 35, 162, 170, 198, 204, 205, 216, 402, 480, 524  
  
 Uhl, Willis L., 545  
 Underhill, O. E., 515  
  
 Vallance, Theodore R., 159  
 Vaughn, K. W., 137, 198  
 Verduin, Jacob, 525  
  
 Vogel, Mabel, 401  
 Votaw, David F., 110, 571  
  
 Walcott, Fred G., 461  
 Walker, Helen M., 338, 370, 390  
 Warner, W. Lloyd, 261  
 Washburne, Carleton, 401  
 Watkins, John G., 545  
 Watson, Goodwin, 31, 32, 35, 222  
 Watson, Richard E., 514  
 Webb, L. W., 18, 36, 64, 198, 277, 338, 391, 418, 461, 480, 503, 525, 545, 581  
 Webb, Sam C., 525  
 Weidemann, Charles C., 159  
 Weitzman, Ellis, 64, 198, 338, 370, 391, 461, 480, 525  
 Wellman, Beth L., 260  
 Wesley, Edgar B., 476, 480  
 Wesman, Alexander G., 255  
 West, Joe Y., 522, 525  
 West, Paul V., 450, 453, 461  
 Whitford, William G., 527, 538, 545  
 Wickman, E. K., 295  
 Wilson, Guy M., 503  
 Wilt, Miriam E., 393  
 Winslow, Leon L., 527, 537, 544  
 Witty, Paul, 418  
 Wood, Ben D., 18, 137, 277, 418, 461  
 Woodruff, Asahel D., 33, 36, 168  
 Woods, Roy C., 545  
 Woodworth, Robert S., 32  
 Woodyard, Ella, 260  
 Workman, Linwood L., 63, 370  
 Wrightstone, J. Wayne, 36, 47, 64, 154, 155, 203, 218, 219, 237, 418, 465, 480  
 Wrinkle, William L., 237  
  
 Yoakam, G. A., 402  
 Young, Robert V., 577, 581  
  
 Zapf, Rosalind M., 521  
 Zimmerman, John G., 514



# Index of Subjects

- Ability, defined, 340
- Accomplishment quotient, 267-68, 346
- Achievement, performance tests of, 52-54
- Achievement quotient, 267-68, 346
- Achievement testing, 13-14
- Adequacy, 75-77
  - defined, 75-76
- Adjustment, 168
- Adjustment inventories, 59, 292-96
- Adjustment measurement, 292-96
- Administrability, 79-80
  - defined, 79
- Age, mental, 257-58
- Age-at-grade norms, 98-99
- Age equivalents, 344
- Age norms, 96, 363
- Alternate-response items, 180-82, 472
  - constructing, 191-92
- Ambiguity, 188
  - freedom from, 90
- American Council on Education Cumulative Record for Elementary and Secondary Schools*, 228
  - illus., Fig. 17, 230-31
- American Council on Education Psychological Examination*, 253, 264
- American Handwriting Scale*, 453
- American School Achievement Tests*, 577
- Analysis (Fig. 1), 50
- Analysis and diagnosis, 113-14
- Analytic tests, 48-49
- Anecdotal method, 283
- Anecdotal record, 61, 297-98
- Answer sheets, separate, 132-34
- Applications, 168, 170, 510
- Appreciations, 168
- Aptitude tests, 31, 56-57, 250-51, 273
- Arithmetic: diagnostic testing in, 491-93
  - organization in, 482-84
  - outcomes in, 484-86
  - remediation in, 494-500
- Arithmetic Achievement Tests*, 490 n.
- Arithmetic mean, defined, 318
- Arithmetic mean of grouped data, computing the, 318
- Arithmetic mean of ungrouped data, computing the, 317
- Arithmetic workbooks, 498-500
- Army Alpha* test, 30, 248
- Army Beta* test, 30, 248, 257
- Army General Classification Test*, 30, 248
- Army Individual Test of Mental Ability*, 30
- "Around the World" Attitudes Inventory, 286
- Art, measurement in, 537-43
- Art appreciation, tests of, 540-42
- Art education: aims of, 537-43
  - outcomes of, 537-39
- Aspects of Personality* inventory: excerpt from, 293
- Table 26, 367

- Association methods, 282  
 Assumed mean, 319  
 Athenians, 22  
 Attitudes, 168, 510-11  
     defined, 285  
     measurement of, 284-87  
 Attitudes scales, 58-59, 285  
*Aviation Cadet Qualifying Examination*, 248  
*Ayres Scale for Measuring the Handwriting of School Children*, 103, 450, 453  
  
 Baker "Telling What I Do" tests, 295  
*Basic Skills, Iowa Every-Pupil Tests of*, cutout scoring stencil, Fig. 5, 130  
     *See also Iowa Every-Pupil Tests of Basic Skills*  
*Basic Writing Vocabulary*, 441  
*Beach Music Test*, 535 n.  
*Beach Standardized Music Test*, 533  
 Behavioral categories (Fig. 22), 302  
*Bell Adjustment Inventory*, 293  
     excerpt from, 60  
*Betts-Keystone Telebinocular*, 411  
*Betts Ready to Read Test*, 405, 411  
 Bi-factor tests, 57, 253-54, 273-74  
*Binet-Simon Scale*, 29-30  
 Bluffing, 165  
 Boston, examinations in, 22  
 Brightness; index of, 263-64  
*Brown-Carlsen Listening Comprehension Test*, 408  
*Buswell-John Diagnostic Chart for Fundamental Processes in Arithmetic*, 492  
  
*California Achievement Tests*, 131, 207, 575-77  
     handwriting scale of (Fig. 12), 208  
*California Arithmetic Test*, profile chart for (Fig. 15), 226  
*California Language Test* (Table 24), 365  
*California Reading Test*, 108  
     *Reading Vocabulary in Social Science*, 473 n.  
*California Short-Form Test of Mental Maturity*, 253  
     excerpts from, 254  
  
*California Test of Mental Maturity*, 253  
 Cardiovascular tests, 556  
 Case study, 297, 298  
 Catch words, 189  
 Central tendency, measures of, 317-28  
 Chance, correction for, 174-75  
 Chance-half coefficient, 74, 386  
 Character Education Inquiry, 32  
 Chart: class analysis, 229-30, 232-33  
     profile, 225-26  
     pupil progress, 226-27, 229  
*Chart for Diagnosing Faults in Handwriting* (Freeman), 454  
 Check lists, 53, 204-6  
*Chicago Tests of Primary Mental Abilities*, 253  
 Chinese examinations, 20  
 City testing bureaus, 122  
 Civics and government, tests in, 470  
*Clapp-Young Self-Marking Tests*, 131  
 Class analysis and diagnosis, 106-7  
 Class analysis chart, 229-30, 232-33  
 Class intervals, 312-16  
 Classroom measurement, practical aspects of, 12-15  
 Classroom testing, 138-39, 160-62  
 Clues and suggestions, 188  
     sparing use of, 89-90  
 Coefficient: chance-half, 74  
     footrule, 74, 387  
     objectivity, 79  
     reliability, 73  
     retesting, 73  
     validity, 70  
*Common School Journal*, 22, 23  
 Comparability, 81-82  
     defined, 81  
*Compass Diagnostic Tests in Arithmetic*, 48  
     excerpt from, 49  
 Completion exercises, 559  
 Completion items, 178-80, 512  
 Composite scores, 353-55  
*Comprehensive Test Program*, 579  
 Computation skills, testing, 489-91  
*Conard Manuscript Writing Standards*, 454  
 Concepts, 168, 169, 510  
 Content areas, achievement batteries in, 578-79  
 Controlled observation, 522  
 Converted scores, 350  
*Cooperative Achievement Tests*, 573-74



- Cooperative General Science Test*, 515 n.  
*Cooperative Mechanics of Expression Test* (Table 25), 366  
*Cooperative Science Test for Grades 7, 8, and 9*, 514 n.  
*Cooperative Social Studies Test for Grades 7, 8, and 9*, 473 n.  
     excerpt from, 184  
 Cooperative Study in General Education, 27, 224, 286, 550-51  
 Cooperative testing programs, 122-23  
*Coordinated Scales of Attainment*, 569-70  
 Correction for chance, 174-75  
 Correlation coefficient, 372-81  
     meaning of, 372, 381-84  
 Counting techniques, 210-13  
 Course objectives, 68-70  
 Criteria, of a good examination, 65-85  
 Cumulative frequency graph, 359-63  
 Curricular validity, 68-70  
 Cutout scoring stencil, *Iowa Every-Pupil Tests of Basic Skills*, Fig. 5, 130  
  
 Deciles, 347  
*Denny-Nelson American History Test*, 474 n.  
 Derived scores, 257-64, 342-55  
     and norms, 342-43  
     based on average performances, 343-44  
     based on variability of performances, 346-50  
     quotients as, 345-46  
 Determining course marks, 195  
 Determining relationships among the test scores, 371-91  
 Diagnosis: Fig. 1, 50  
     meaning and importance of, 113-16  
     nature of, 114-15  
 Diagnostic and analytic tests, 48  
 Diagnostic profile charts, 108, 110  
 Diagnostic testing, 12-14  
 Diagnostic tests, 48-49  
     of problem-solving ability, 494  
 Differential aptitudes tests, 254, 274  
     excerpts from, 255  
 Difficulty, 90-91  
 Direct observation, 61, 281, 300  
 Directed observation, 283  
 Discriminative power, 91-93  
  
 Dispersion, measures of, 328-37  
 Double-entry table, 374-75  
 Double negatives, 191  
*Drake Musical Memory Test*, 532  
 Drawing scales and tests, 540  
 Drill, 117-18  
 Drill tests, 50  
 Duplicate forms, of test, 82  
*Durrell Analysis of Reading Difficulty*, 411  
*Durrell-Sullivan Reading Capacity Tests*, 405  
*Dyer Backboard Test of Tennis Ability*, 559  
*Dyer-Schurig-Apgar Basketball Test*, 559  
  
 Economy, defined, 81  
 Economy of time, 165  
 Educational and mental tests, 37-38  
 Educational evaluation. *See* Evaluation  
 Educational measurement: first book on, 24  
     present status of, 33-34  
     *See also* Measurement  
 Educational quotient, 345  
 Educational testing: from 1800 to 1900, 22-24  
     from 1900, 24-27  
 Educational tests, 38, 44-54  
     described, 5-6  
     early, 21-22  
     general characteristics of, 5-8  
 Eight-Year Study, 27  
 Elementary mathematics, measurement and evaluation in, 481-503  
 Elementary science: diagnosis and remediation in, 522-23  
     measurement and evaluation in, 504-25  
     objectives of, 505-7  
     outcomes of, 507-11  
     testing in, 514-18  
     tests in, 511-16  
*Ely-King Tests in American History*, 472 n.  
 Emotional adjustment, measurement of, 291-96  
 Ephraimites, 20  
 Equated scores, 350  
 Error of grouping, 312  
 Essay examinations, 44, 45  
 Essay questions, types of, 151-53

- Essay tests, 42-43, 141-57  
   advantages of, 147-50  
   conclusions concerning, 150-51  
   improving, 151-57  
   limitations of, 142-47  
   scoring, 153-55  
   suggestions for improving, 156-57  
 Establishing reliability, 103-5  
 Establishing validity, 103-5  
 Evaluation, 3-4  
   in elementary mathematics, 481-503  
   in elementary sciences, 504-25  
   in expressive language arts, 419-61  
   in fine arts, 526-45  
   in health and physical education, 544-63  
   in receptive language arts, 393-418  
   in social studies, 462-80  
   meaning of, 218-19  
   need for, in education, 1-3  
   of general educational achievement, 564-81  
   *See also* Measurement  
 Evaluation tools and techniques, 217-37  
 Evaluations, personality, 58-61  
 Evaluative instruments, 3  
   and techniques, 44  
   development of, 27  
 Evaluative techniques, 27, 54, 61, 232-35, 297-303  
 Evaluative tests, 53, 219-25  
 Evaluative tools, 27, 54, 225-33  
   using, 234-35  
 Examination, characteristics of a good, 65-85  
 Examinations, in Boston, 22  
 Exercises, matching, 185-86  
   constructing, 193-94  
 Expectancy, standards of (Table 7), 145  
 Expressive language arts, measurement and evaluation in, 419-61  
 Extensive sampling, 163-64  
   effect of (Fig. 10), 164  
  
 Factor analysis, 31  
 Faculty theory, 241  
 Fine arts, measurement and evaluation in, 526-45  
 Footrule coefficient, 74, 387  
*Franseen Diagnostic Tests in Language*, 433  
 Free association, 281  
  
*Freeman Chart for Diagnosing Faults in Handwriting*, 454  
*French-Cooper Volleyball Test*, 559  
 Frequency polygon, 356-58  
 Frequency table, 313  
  
*Gates Primary Reading Tests*, 410  
*Gates Reading Readiness Test*, 404  
*Gates Silent Reading Tests*, 409  
*Gates-Strang Health Knowledge Test*, 549  
 General achievement, measurement of, 564-67  
 General achievement batteries, 567-74  
 General educational achievement, measurement and evaluation in, 564-81  
 General intelligence, 248-49  
   individual scales of, 244-48  
 General intelligence tests, 55-56, 244-49, 270-72  
 General science, 508-9  
*Generalized Attitudes Scales*, 285  
 Geography, tests in, 470  
*Gettysburg Edition*, 453  
 Gileadites, 20  
*Gilmore Oral Reading Test*, 405-6  
   excerpt from, 406  
 Grade equivalents, 343-44  
 Grade norms, 95-96, 363  
 Grammar and usage, standardized tests in, 432-33  
 Graphical representation, 355-63  
*Gray's Oral Reading Check Tests*, 406, 407  
*Gray Standardized Oral Reading Paragraphs*, 405, 406  
*Gray-Votaw-Rogers General Achievement Tests*, 108, 571-72  
   Fig. 4, 110  
 Greenwich Astronomical Observatory, 28  
 Greenwich Hospital School, 23  
 Group comparisons, 112  
 Group data, 309-16  
 Group dynamics, evaluation of, 61, 299-303  
 Group factors, 241  
 Group-factor tests, 57, 253-55, 273-74  
 Group intelligence tests, 55-56  
 Grouping data, 309-16  
*Guess Who Test*, 293-94  
 Guessing, 166-67



- Haggerty-Olson-Wickman Behavior Rating Schedules*, 294  
 excerpt from, 295  
 Handedness in writing, 457-58  
 Handwriting, measurement and remediation of, 448-57  
 Handwriting scale, *California Achievement Tests* (Fig. 12), 208  
*Harrison-Stroud Reading Readiness Tests*, 404  
*Hayes Scale for Evaluating the School Behavior of Pupils*, 296  
*Health Activities Inventory*, 223, 234  
 excerpts from, 224  
 Health and physical education, measurement and evaluation in, 544-63  
 Health attitudes inventories, 550  
*Health Attitudes Inventory*, 286, 288, 550  
*Health Awareness Test*, 550  
 excerpts from, 551  
 Health education: measurement in, 544-53  
 prevention and diagnosis in, 552-53  
 scope and aims of, 544-48  
 Health evaluation inventories, 550-51  
*Health Interests Inventory*, 288  
 Health inventories, 223-24, 552  
 Health knowledge tests, 548-50  
*Heath-Rodgers Soccer Test*, 559  
*Hillegas Composition Scale*, 432  
 Histogram, 358-59  
 History, tests in, 469
- Index of brightness, 263-64  
 Index of studiousness, 268  
 Individual behavior, evaluation of, 61, 297-99  
 Individual differences, recognition of, 21, 28  
 Individual intelligence scales, 55  
 Individual pupil diagnosis, 107-8  
 Informal objective testing, using results, 194-95  
 Informal objective tests, 4, 9, 43, 160-98  
 advantages of, 163-65  
 construction of, 167-73  
 development of, 25-26  
 possible disadvantages of, 165-67  
 using, 173-75  
 Instructional outcomes, 68-70  
 types of, 168-70
- Instructional tests, 50  
 Integral limits, 313-16  
 Intelligence, 13  
 bi-factor tests of, 253-54  
 defined, 239  
 early attempts to measure, 29  
 group-factor tests of, 253-55, 273-74  
 measurement of, 242-49  
 multi-factor tests of, 253-55  
 nature of, 239-41  
 performance tests of, 57-58, 274  
 Intelligence and aptitude tests, 238-76  
 Intelligence quotient, 258-62, 345  
 constancy of, 260-61  
 distribution of the, 265-66  
 future of the, 262  
 social class and the, 261-62  
 Intelligence testing, 13  
 derived results of, 257-64  
 from 1800 to 1900, 28-29  
 from 1900 to the present, 29-31  
 procedures for, 268-70  
 Intelligence tests, 38, 54-58  
 administering and scoring, 269  
 factored, 31  
 first individual, 29-30  
 group, 30  
 individual, in America, 30  
 specific, 31, 273  
 uses of, 269-74  
*Interest-Attitude Test* (Pressey), 288  
 Interests, 168, 510-11  
 defined, 287  
 informal measurement of, 289  
 measurement of, 287-90  
 Interests inventories, 59, 287-90  
 Interests measurement, 287-90  
 International Test Scoring Machine, 131  
 Fig. 6, 132  
*Interpretation of Data Test*, 220-21, 235  
 excerpt from, 221  
 Interpreting test results, 339-70  
 Interpretive tests, 53-54, 220-23  
*Inter-Trait Rating Scale*, 303  
 Interview, 233-34, 287  
 Inventories: adjustment, 292-96  
 interests, 287-90  
 personality, 58-61  
*Inventory of Personal-Social Relationships*, 224  
 excerpt from, 225  
 Inventory tests, 46, 47

- Iowa Algebra Aptitude Test*, 250, 500  
 excerpt from, 251
- Iowa-Brace Test*, 556
- Iowa Elementary Language Tests*, 434
- Iowa Every-Pupil Tests of Basic Skills*,  
 434, 490 n., 574-75  
 cutout scoring stencil, Fig. 5, 130  
 Language, 40, 98, 100  
 excerpts from, 39  
 Table 4, 99  
 Table 5, 101
- Iowa General Information Test in American History*, 472 n.
- Iowa Grammar Information Test*, 433
- Iowa Language Abilities Test*, 95-96,  
 433, 435  
 excerpts from, 434  
 Table 3, 97
- Iowa Placement Examinations*, 250
- Iowa Revision of the Brace Test of Motor Ability*, 555-56
- Iowa Silent Reading Tests*, 409
- Iowa Spelling Scales*, 51, 441, 442
- Items: alternate-response, 180-82, 191-92  
 completion, 178-80  
 constructing, 189-93  
 multiple-choice, 182-84, 192-93  
 multiple-response, 183-84  
 recall-type, 189-91  
 simple recall, 177-78
- Kellogg-Morton Revised Beta Examination*, 257
- Kirby Grammar Test*, 432
- Kline-Carey Drawing Scales*, 540
- Knauber Test of Art Ability*, 542
- Knowledges, 168, 169, 510
- Kuhlmann-Anderson Intelligence Tests*,  
 248, 263  
 excerpts from, 249
- Kwakwasser-Dykema Music Tests*, 531
- Kwakwasser-Ruch Test of Musical Accomplishment*, 535 n.
- Kwakwasser Test of Music Information and Appreciation*, 534 n., 536
- Language: importance of, 419-20  
 remedial instruction in, 434-39
- Language, *Iowa Basic Skills Tests*, 98  
 Table 4, 99  
 Table 5, 101
- Language abilities, measurement of,  
 433-34
- Language arts: expressive, measurement  
 and evaluation in, 419-61  
 receptive, measurement and evaluation in, 393-418
- Language skills, analysis of, 420-24
- Learning, measuring efficiency of, 112
- Lee-Clark Reading Readiness Test*, 405
- Letter marks, 352-55
- Lewerenz Test in Fundamental Abilities of Visual Arts*, 542
- Limited sampling, 77, 142-43
- Listening and reading, significance of,  
 393-94
- Listening comprehension, measurement  
 of, 407-8
- Listening efficiency, 399
- Logical Reasoning Test*, 222, 235  
 excerpt from, 223
- Logical validity, 72
- Maladjustment, 291
- Manuscript Writing Standards (Conard)*, 454
- Master Achievement Tests*, 572
- Mastery tests, 50
- Matching exercises, 185-86, 474-75, 514-15, 558-59  
 constructing, 193-94
- Mathematics: elementary, measurement  
 and evaluation in, 481-503  
 prediction of success in, 500-501
- McAdory Art Test*, 540, 541
- McCall Inter-Trait Rating Scale*, 304
- McCauley Examination in Public School Music*, 534 n.
- Measuring, 3
- Measuring techniques, 210-13
- Measurement: and the total child, 14-15  
 development of, 19-36  
 general vs. specific, 564-65  
 in elementary mathematics, 481-503  
 in elementary sciences, 504-25  
 in expressive language arts, 419-61  
 in fine arts, 526-45  
 in health and physical education, 544-63  
 in receptive language arts, 393-418  
 in social studies, 462-80  
 need for in education, 1-3



- Measurement (*cont.*)  
 of general educational achievement,  
   564-81  
 standard error of, 75, 387-88  
 to 1800, 20-22  
*See also* Evaluation
- Median, defined, 324
- Median of grouped data, computing  
 the, 324
- Meier Art Judgment Test*, 540  
 excerpts from, 541
- Meier-Seashore Art Judgment Test*, 541
- Mental age, 257-58
- Mental growth, 259
- Mental measurement, present status of,  
 33-34
- Mental Measurements Yearbooks*, 123,  
 125
- Merrill-Palmer Scale of Mental Tests*,  
 264
- Metronoscope, 411
- Metropolitan Achievement Tests*, 229,  
 474 n., 475 n., 568-69  
 Fig. 16, 227
- Metropolitan Literature Test*, excerpt  
 from, 186
- Metropolitan Readiness Test*, 252, 405
- Metropolitan Reading Test*, excerpt  
 from, 180
- Mid-measure, 323
- Midpoint, 313-16
- Modern School Achievement Tests*, 511,  
 570-71, 578
- Monroe Reading Aptitude Tests*, 405
- Mooney Problem Check List*, 294
- Multi-factor tests, 57, 253-55, 274
- Multi-factor theory, 241
- Multiple-attribute tests, 41
- Multiple-choice items, 182-84, 473-74,  
 489-91, 513-14, 558  
 constructing, 192-93
- Multiple-response items, 183
- Murphy-Durrell Diagnostic Reading  
 Readiness Test*, 405, 408
- Music, measurement in, 527-37
- Music education, aims and outcomes of,  
 527-30
- Musical achievement, measurement of,  
 533-37
- Musical Achievement Test*, 535 n.
- Musical memory, measurement of, 532-  
 33
- Musical talent, measurement of, 530-33
- National Achievement American History  
 Test*, excerpt from, 178
- National Achievement Tests*, 573-74
- Nature study, 508
- Navy General Classification Test*, 248
- New Revised Stanford-Binet Tests of  
 Intelligence*, 245-48, 264
- Non-test tools, 43-44
- Non-verbal tests, 42
- Norms, 82, 342-43, 363-67  
 deriving test, 94-102  
 reliability of, 104-5  
 types of, 95-102
- Norms vs. standards, 102-3
- Object tests, 52, 201, 202-4
- Objective examinations and scales, 44,  
 45-51
- Objective test items, constructing, 186-  
 94
- Objective tests, 38, 43  
 early, 23-24  
 functions of, 6-8  
 standardized vs. non-standardized,  
 161-62
- Objectives, reading and listening, 395-98
- Objectives of the social studies, 463-64
- Objectivity, 4, 77-79, 89, 389  
 coefficient of, 79  
 defined, 77  
 determination of test, 389
- Observational methods, 283
- Ophthalmograph, 411
- Oral examinations, 44-45  
 early, 20
- Oral language, diagnosis of, 426-28
- Oral language scales, 425-26
- Oral language skills, 424-25
- Oral reading, remedial drills in, 411-13
- Oral reading check tests, 406-7
- Oral reading paragraphs, 405-6
- Oral tests, 42, 138-41  
 advantages of, 140-41  
 limitations of, 140
- Orleans Algebra Prognosis Test*, 500
- Otis Classification Test*, 579
- Otis Quick-Scoring Group Tests of  
 Mental Ability*, 264
- Outcomes: in arithmetic, 484-86  
 of elementary science, 507-11  
 of social studies, 465-67  
 reading and listening, 395-98

- Pearson product-moment coefficient of correlation, 372-81
- Per cent of average development, 263
- Percentile, defined, 347
- Percentile grade norms, 99-100
- Percentile norms, 363  
for school averages, 100-102
- Percentile ranks, 346-47, 362  
defined, 347
- Percentile scores, 264
- Percentiles, 347-48, 362
- Performance, defined, 340
- Performance measures, 201, 204-7
- Performance testing, using results of, 215
- Performance tests, 42, 52-54, 57-58, 199-216  
and scales, 44  
constructing, 213-14  
nature of, 200-1  
of intelligence, 274  
and aptitude, 256-57
- Personal constant, 263
- Personal Data Sheet*, 32
- Personal reports, 281, 284  
blanks, 292-94
- Personality: defined, 280  
measurement of total, 303-5  
nature of, 279-81
- Personality evaluation, 58-61  
1800 to the present, 32-33
- Personality instruments and techniques, 278-305
- Personality inventories, 32, 58-61
- Personality measurement: techniques of, 281-84  
tools of, 284-96
- Personality quotient, 303-5
- Personality testing, 13
- Personality tests, 38  
antecedents of modern, 32
- Personality types, classification of, 21
- Physical classification tests, 559-60
- Physical education: diagnosis in, 560  
measurement in, 553-60  
objectives of, 553-54  
tests in, 557-59
- Physical education and health, measurement and evaluation in, 544-63
- Physical examinations, 551-52
- Physical qualities, tests of, 555-56
- Physiology and hygiene, 508
- Pintner General Ability Tests*, 253  
excerpts from, 56  
Table 23, 364
- Pintner-Paterson "Long" Performance Scale*, 257  
Fig. 19, 256
- Posture tests, 556
- Power test, 40
- Practicality, 79-81
- Practice exercises, 50, 117
- Practice tests, 50
- Practices and activities, tests of, 223-25
- Prediction, significance of correlation coefficient for, 382-84
- Pressey Interest-Attitude Test*, 59, 288  
excerpts from, 60
- Preventive work, 116
- Primary mental abilities tests, 274
- Primitive tribes, 21
- Problem-solving, testing, 493-94
- Problem-solving exercises, 495-98
- Product evaluation, 201, 207-13
- Product moments, 377-79
- Product scales, 51
- Profile chart, 225-26  
for *California Arithmetic Test* (Fig. 15), 226
- Prognostic Test of Mechanical Abilities*, 47, 202  
excerpts from, 203
- Prognostic tests, 46, 47
- Progress chart, 226-27, 229
- Progressive Achievement Tests*, 575-76
- Progressive Tests in Social and Related Sciences*, 472 n., 511, 513 n., 578-79  
excerpts from, 182, 550
- Projective method, 32-33, 61
- Projective techniques, 297, 298-99
- Psychological and logical validity, 68
- Psychological examinations, 38
- Psychological validity, 72
- Pupil gradation and placement, 111-12
- Pupil record, cumulative, 228-29
- Quality scale, 52, 207
- Quartiles, 347
- Questionnaire, 234
- Questions, discussion, short answer, simple-recall, 153
- Quintiles, 347



- Quizzes, 50
- Quotient: accomplishment, 267-68,  
346  
achievement, 267-68, 346  
educational, 345  
intelligence, 258-62, 345  
personality, 303-5  
reading, 345
- Range, 312, 313  
defined, 330
- Rapport, 245
- Rating Form for Fastening, excerpt  
from, 209
- Rating scales, 52-53, 207-8, 281, 283-84,  
294-96
- Readability, 401-2
- Read General Science Test*, 513 n.  
excerpt from, 184
- Readiness, reading, 402-5
- Readiness tests, 57, 251-52, 273
- Reading: and listening, objectives and  
outcomes of, 395-98  
corrective exercises in, 411-15  
defects in, 400  
oral vs. silent, 400-401
- Reading quotient, 345
- Reading readiness, 402-5  
tests, 403-5
- Real limits, 313-16
- Recall items, 46  
constructing, 189-91
- Receptive language arts, measurement  
and evaluation in, 393-418
- Recognition items, 46
- Relationships, measures of, 371-91
- Relative ranks, 351-52
- Reliability, 4, 72-79, 385-88  
defined, 72  
establishing, 103-5  
evaluation of test, 385-88
- Reliability coefficient, 73, 385
- Remedial drill, 117
- Remedial teaching, 12-14
- Report card, pupil, 229
- Response, uniformity of, 89
- Retesting coefficient, 73, 385
- Rogers Personality Test*, 292
- Rorschach test*, 32, 298
- Rounding numbers, 315
- Ruch-Popenoe General Science Test*,  
512 n.
- Russell-Lange Volleyball Test*, 559
- Sampling: intensive, 142-43  
principle of (Fig. 2), 76
- Scale books, 23-24
- Scale for Handwriting of Children*, 25
- Scale for Measuring the Handwriting of  
School Children* (Ayres), 103
- Scaled scores, 350
- Scaled tests, 38-40
- Scales, 38-39  
attitudes, 285  
oral language, 425-26
- Scales for the Measurement of Social  
Attitudes* (Thurstone), 285
- Scaling, 39
- Scatter diagram, 374-75
- Sciences, elementary, measurement and  
evaluation in, 504-25
- Scientific attitude, measurement of, 519-  
21
- Scorability, 80-81
- Score card, 52-53, 207, 209
- Score Card for Waffles, 209
- Scores: composite, 353-55  
derived, 257-64  
percentile, 264  
standard, 264
- Scoreze*, 131
- Scoring: objectivity of, 165  
subjectivity of, 143-45
- Scoring machines, 131
- Seashore Measures of Musical Talent*,  
530, 531
- Simple recall items, 177-78, 471, 489,  
490, 512
- Skill areas, achievement batteries in,  
574-78
- Skills, 168, 169, 510  
basic arithmetic, 486-88  
written language, 428-31
- Social education, 462-63
- Social learning, 462
- Social studies: defined, 462  
evaluation in, 476-77  
informal objective tests in, 475  
measurement and evaluation in, 462-  
80  
objectives of, 463-64  
organization of, 464-65  
outcomes of, 465-67  
remediation in, 477-78
- Social studies tests, 467-75
- Sociogram, 61, 299-301  
Fig. 21, 301

- Sociometric method, 299-303
- Source scales, 51
- Spartans, 21
- Spearman-Brown Prophecy Formula*, 74, 386
- Specialized achievement batteries, 579
- Specific determiners, 90, 191
- Specific intelligence tests, 56-57, 250-52
- Speech disorders, 427-28
- Speed test, 40
- Spelling: diagnosis and remediation in, 444-48
  - measurement and remediation in, 439-48
  - objectives of, 439-40
- Spelling tests, 440-43
- SRA Primary Mental Abilities Tests*, 253
- Standard deviation, 330-37
  - defined, 330
  - derived scores based on the, 348-50
  - of grouped data, computing the, 333-37
  - of ungrouped data, computing the, 332-33
- Standard Elementary Spelling Scale*, 442
- Standard error of measurement, 75, 387-88
- Standard measures, 349
- Standard scores, 264, 350
- Standardization, meaning of, 87
- Standardized achievement tests: first, 25
  - later development of, 26-27
- Standardized educational tests, 4, 9-12
- Standardized tests, 42, 43, 86-137
  - administration of, 11
  - administrative uses of, 109-12
  - constructing, 86-105
  - guidance uses of, 109
  - instructional uses of, 105-8
  - practical uses of, 105-19
  - scoring of, 11-12
  - selection of, 11
- Stanford Achievement Test*, 433, 511, 567-68
- Science*, 513 n.
- Stanford Revision of the Binet Scale*, 30, 246
- Statistical methods, 308-89
  - foundations of, 28
- Statistical validity, 68, 70-72
- Strong Vocational Interest Blank*, 288
- Studiosness, index of, 268
- Subject norms, 99-100
- Summarizing test results, 308-38
- Superstitious beliefs, measurement of, 521-22
- Survey tests, 46
- Szondi Test*, 298
- Tabulation of test scores, 309-16
- Tachistoscope, 411
- Tastes and preferences, 168
- Teacher-made tests, 9, 42-43
- Teacher's Word Book*, 441
- Techniques, 43-44
  - evaluative, 232-35, 297
  - projective, 298-99
  - verbal association, 282
  - visual stimulus, 282
- "*Telling What I Do*" tests (Baker), 295
- Terman Group Test of Mental Ability*, 107
- Test, 38
  - reliability of, 104
  - validity of, 104
- Test check list, 125
- Test content, validity of, 87-88
- Test forms, equating, 93-94
- Test items, constructing and validating, 88-93
- Test norms, 363-67
- Test rating scales, 124-25
- Test results, analyzing and interpreting, 12, 134
- Test score: defined, 340
  - meaning of a, 315, 339-42
  - tabulation of, 309-16
- Test Score Card, 157
- Testing, 3
- Testing programs: nation-wide, 123
  - planning, 119-23
  - state-wide, 122-23
  - steps in, 119
- Tests, 38-40, 43-44
  - administering, 126-28
  - aptitude, 250-51, 273
  - bi-factor, 273-74
  - essay, 141-57
  - evaluative, 219-25
  - first in the school, 22
  - general classification of, 37-44
  - general intelligence, 244-49, 270-72



Tests (*cont.*)

- hand-scored, 129-31
- in elementary science, 511-16
- informal objective, 160-98
- intelligence and aptitude, 238-76
- interpretive, 220-23
- inventory, 46, 47
- machine-scored, 131
- multi-factor, 274
- non-verbal, 41-42
- of practices and activities, 54, 223-25
- of proficiency in sports, 559
- oral, 138-41
- performance, 41-42, 199-216
- power, 40
- prognostic, 46, 47
- readiness, 251-52, 273
- reading readiness, 403-5
- scoring, 128-33
- selecting, 123-26
- self-scoring, 131
- significance of, 8-9
- specific intelligence, 250-52, 273
- speed, 40
- spelling, 440-43
- standardized, 42, 43, 86-137
- survey, 46
- teacher-made, 42-43
- types of, 8-12, 37-64
- verbal, 41-42
- when to give, 121-22
- Thematic Apperception Test*, 298
- Thorndike Extension to the Hillegas Scale for the Measurement of Quality in English Composition by Young People*, 432
- Thorndike Scale*, 453
- Thurstone Scales for the Measurement of Social Attitudes*, 285
- Time-limit test, 41
- Timing devices, 206-7
- Tools, evaluative, 225-33
- Total child, measurement and the, 14-15
- True-false items, 512-13, 557

T-scores, 349-50

Two-factor theory, 241

Understandings, 168, 169-70, 510

*Unit Scales of Attainment*, 569*Unit Scales of Attainment in Language*, 433

University of Bologna, 22

University of Paris, 22

Using the informal objective test, 173-75

Utility, 82-83

Validity, 4, 66-72, 385

coefficient of, 70

curricular, 68-70

defined, 66

determination of test, 385

establishing, 103-5

*Van Wagenen Reading Readiness Test*, 405

Variability, measures of, 328-37

Verbal association techniques, 282

Verbal tests, 41-42

Visual stimulus techniques, 282

*Watson-Glaser Test of Critical Thinking*, 221

excerpts from, 222

Whole child, 3-4

*Willing Scale for Measuring Written Composition*, 432

Work-limit test, 40

Work-study reading: measurement of, 408-10

remedial drills in, 413-15

Written composition, measurement of, 432-33

Written examinations, early, 20

Written language skills, 425, 428-31

Written quiz, 45

Z-scores, 349

Form No. 3.

PSY, RES.L-1

**Bureau of Educational & Psychological  
Research Library.**

The book is to be returned within  
the date stamped last.

21.6.6

23.6.60

20 FEB 1961

25 APR 1963

16.8.63

12.7.65

30.6.66

16.8.67

27.4.69

5.1.76